# NAVAL POSTGRADUATE SCHOOL

**MONTEREY, CALIFORNIA**

# THESIS

**FORECASTING U. S MARINE CORPS REENLISTMENTS BY MILITARY OCCUPATIONAL SPECIALTY AND GRADE**

by

Dean G. Conatser

September 2006

| | |
|---|---|
| Thesis Advisor: | Ronald D. Fricker, Jr. |
| Second Reader: | Samuel E. Buttrey |

**Approved for public release; distribution is unlimited**

THIS PAGE INTENTIONALLY LEFT BLANK

**REPORT DOCUMENTATION PAGE**

*Form Approved OMB No. 0704-0188*

**1. AGENCY USE ONLY *(Leave blank)***

**2. REPORT DATE**
September 2006

**3. REPORT TYPE AND DATES COVERED**
Master's Thesis

**4. TITLE AND SUBTITLE** Forecasting U.S. Marine Corps Reenlistments by Military Occupational Specialty and Grade

**5. FUNDING NUMBERS**

**6. AUTHOR** Dean G. Conatser

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Naval Postgraduate School
Monterey, CA 93943-5000

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
N/A

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**

**11. SUPPLEMENTARY NOTES** The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited

**12b. DISTRIBUTION CODE**
A

**13. ABSTRACT**

Each year, manpower planners at Headquarters Marine Corps must forecast the enlisted force structure in order to properly shape it according to a goal, or target force structure. Currently the First Term Alignment Plan (FTAP) Model and Subsequent Term Alignment Plan (STAP) Model are used to determine the number of required reenlistments by Marine military occupational specialty (MOS) and grade. By request of Headquarters Marine Corps, Manpower and Reserve Affairs, this thesis and another, by Captain J.D. Raymond, begin the effort to create one forecasting model that will eventually perform the functions of both the FTAP and STAP models.

This thesis predicts the number of reenlistments for first- and subsequent-term Marines using data from the Marine Corps' Total Force Data Warehouse (TFDW). Demographic and service-related variables from fiscal year 2004 were used to create logistic regression models for the FY2005 first-term and subsequent-term reenlistment populations. Classification trees were grown to assist in variable selection and modification. Logistic regression models were compared based on overall fit of the predictions to the FY2005 data.

Combined with other research, this thesis can provide Marine manpower planners a means to forecast future force structure by MOS and grade.

**14. SUBJECT TERMS** Reenlistments, Marine Corps Manpower, Total Force Data Warehouse

**15. NUMBER OF PAGES**
92

**16. PRICE CODE**

| | | | |
|---|---|---|---|
| **17. SECURITY CLASSIFICATION OF REPORT**<br>Unclassified | **18. SECURITY CLASSIFICATION OF THIS PAGE**<br>Unclassified | **19. SECURITY CLASSIFICATION OF ABSTRACT**<br>Unclassified | **20. LIMITATION OF ABSTRACT**<br>UL |

THIS PAGE INTENTIONALLY LEFT BLANK

# FORECASTING U. S. MARINE CORPS REENLISTMENTS BY MILITARY OCCUPATIONAL SPECIALTY AND GRADE

Dean G. Conatser
Major, United States Marine Corps
B.S., United States Air Force Academy, 1994

Submitted in partial fulfillment of the
requirements for the degree of

**MASTER OF SCIENCE IN OPERATIONS RESEARCH**

from the

**NAVAL POSTGRADUATE SCHOOL**
**September 2006**

Author: Dean G. Conatser

Approved by: Ronald D. Fricker, Jr.
Thesis Advisor

Samuel E. Buttrey
Second Reader

James N. Eagle
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

# ABSTRACT

Each year, manpower planners at Headquarters Marine Corps must forecast the enlisted force structure in order to properly shape it according to a goal, or target force structure. Currently the First Term Alignment Plan (FTAP) Model and Subsequent Term Alignment Plan (STAP) models are used to determine the number of required reenlistments by Marine military occupational specialty (MOS) and grade. By request of Headquarters Marine Corps, Manpower and Reserve Affairs, this thesis and another, by Captain J.D. Raymond (Raymond, 2006), begin the effort to create one forecasting model that will eventually perform the functions of both the FTAP and STAP models.

This thesis predicts the number of reenlistments for first and subsequent-term Marines using data from the Marine Corps' Total Force Data Warehouse (TFDW). Demographic and service-related variables from fiscal year 2004 were used to create logistic regression models for the FY2005 first-term and subsequent-term reenlistment populations. Classification trees were grown to assist in variable selection and modification. Logistic regression models were compared based on overall fit of the predictions to the FY2005 data.

Combined with other research, this thesis can provide Marine manpower planners a means to forecast future force structure by MOS and grade.

THIS PAGE INTENTIONALLY LEFT BLANK

# TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| AFQT | Armed Forces Qualification Test |
| ASR | Authorized Strength Report |
| CFRM | Career Force Retention Model |
| CNA | Center for Naval Analyses |
| DOD | Department of Defense |
| ECC | End of Current Contract |
| FTAP | First Term Alignment Plan |
| FY | Fiscal Year |
| GAO | U.S. Government Accountability Office |
| GAR | Grade Adjusted Recapitulation |
| LDS | Longitudinal Data Set |
| M&RA | Manpower and Reserve Affairs |
| MCCDC | Marine Corps Combat Development Command |
| MCRC | Marine Corps Recruiting Command |
| MOS | Military Occupational Specialty |
| NPS | Naval Postgraduate School |
| Occfield | Occupational Field |
| P2T2 | Patient, Prisoner, Trainee, and Transient |
| SRB | Selective Reenlistment Bonus |
| SSN | Social Security Number |
| STAP | Subsequent Term Alignment Plan |
| TECOM | Training and Education |
| TFSD | Total Force Structure Division |
| USMC | United States Marine Corps |
| YOS | Years of Service |

THIS PAGE INTENTIONALLY LEFT BLANK

# ACKNOWLEDGMENTS

I would like to thank Dr. Ron Fricker and Dr. Sam Buttrey for their patience and instruction.  I greatly appreciate that they shared their time and knowledge with me.

To my family, thank you for your love and encouragement.  Finally, to my friends, thank you for your great input.

THIS PAGE INTENTIONALLY LEFT BLANK

# EXECUTIVE SUMMARY

The U. S. Marine Corps currently uses three models to forecast the enlisted force structure each year. Two of the models, the First Term Alignment Plan (FTAP) model, and the Subsequent Term Alignment Plan (STAP) model are used to determine the number of reenlistments required to meet future force goals, as depicted in the Grade Adjusted Recapitulation (GAR). Such planning tools are essential in managing an enlisted force of roughly 160,000 enlisted Marines. At the request of Headquarters Marine Corps, Manpower and Reserve Affairs, this work was begun to explore the possibility of creating a single model to perform the functions of both the FTAP and STAP models. When completed, the new model will be called the Career Force Retention Model (CFRM).

The purpose of this thesis is to predict the number of reenlistments for both first-term Marines and subsequent-term Marines, by military occupational specialty (MOS) and grade. Combining the output of reenlistment forecasts with predictions made on the population of Marines not approaching the end of enlistment will result in a forecast of the overall force structure. This thesis and Captain J.D. Raymond's thesis, entitled *Determining the Number of Reenlistments Necessary to Satisfy Future Force Requirements* (Raymond, 2006), are the beginning of the development of the CFRM.

Using data from the Marine Corps' Total Force Data Warehouse (TFDW), a longitudinal data set was formed to utilize demographic and service-related variables for Marines with contracts ending in FY2004 and FY2005. SRB multiples offered to the individual Marines' MOS and SRB Zone were merged with the TFDW data. Demographic variables included AFQT score, number of dependents, race and ethnicity, marital status, and gender. Service-related variables included grade, SRB multiple offered to reenlist, MOS, and years of service.

Since no data explicitly identified Marines as having extended, Marines who reenlisted or extended a current enlistment contract were treated alike. That is, both Marines who reenlisted and Marines who extended their contract from one year to the

next were indistinguishable (in the available data) and thus the combined groups were simply classified as having been "retained." Marines with contracts ending in FY2004 were grouped to create a model for predicting FY2005 reenlistments. This was done for both the first- and subsequent-term populations.

Two classification trees were made on the FY2004 reenlistment data in order to develop a working knowledge of which variables would likely be most important in forecasting retention. The structure of the trees indicated that the Marines' grade and years of service were useful in prediction. After cross-validation and pruning, the trees did not achieve better than 70 percent correct classification for first-term Marines and 75 percent for subsequent-term Marines. However, the trees did provide useful information on which variables might be the most useful in predictions by logistic regression. Further, the levels at which the trees split the variables offered insight into how categorical variables might be collapsed, or numeric variables modified to be categorical.

To predict the total number of expected FY2005 reenlistments by MOS and grade, logistic regression models were created for the first and subsequent-term populations. By using a chi-square-like statistic to measure overall goodness of fit, the models were compared, and "winners" chosen. The best models for both populations were very similar, using grade, years of service, and ethnicity as predictors. Differences in the variables used existed only in the modifications to their raw form, as suggested by the classification tree splits. The table below provides an example of predictions for the FY2005 first-term population having MOS 0311 and GRADE E-4. In Table 1 below, the predicted number of reenlistments is compared to the actual, with measures of error to the right.

Table 1. Comparison of reenlistment predictions for E-4s in MOS 0311.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 1590 | 556.97 | 466.00 | 90.97 | 8275.48 | 14.86 |
| B | 1590 | 505.90 | 466.00 | 39.90 | 1591.84 | 3.15 |
| M | 1590 | 489.19 | 466.00 | 23.19 | 537.91 | 1.10 |

For the first-term population, model “M” (defined in Chapter V) was overall the best model, as determined by goodness-of-fit over all 726 MOS and grade combinations. For both the first and subsequent terms, the best model did not dominate across all individual MOSs and grades.

Surprisingly, SRB multiple offered for reenlistment was not a strong predictor in logistic regression. This result was foreshadowed by the classification trees’ omission of the SRB variable altogether. The lack of contribution by the SRB data in reenlistment prediction suggests that further research is warranted in determining SRB allocation. Other future work in this area should include deployment data from TFDW. A variable accounting for deployed time, especially given today’s high operational tempo, could be valuable in forecasting reenlistment.

THIS PAGE INTENTIONALLY LEFT BLANK

# I. INTRODUCTION

## A. PURPOSE

Each year the Marine Corps must determine the number of reenlistments required to meet its force requirements. The purpose of this thesis is to provide manpower planners at Headquarters Marine Corps with a tool to forecast the number of reenlistments by military occupational specialty (MOS) and pay grade. At the request of the Manpower and Reserve Affairs (M&RA) department of Headquarters Marine Corps, the output of this thesis will be integrated with another thesis that calculates the force distribution of Marines who are not approaching the end of their contracts.

With the forecasts of these two models combined, Marine Corps manpower planners can determine which categories, indexed by MOS and grade, are likely to be under and over their acceptable manning levels in the next year. Such a forecast is required for planners to estimate the number of required new recruits and to effectively utilize such measures such as the Selective Reenlistment Bonus (SRB) to influence the retention of enlisted Marines.

## B. BACKGROUND

### 1. Brief Overview of the Manpower Planning Process

The Deputy Commandant of the Marine Corps for Manpower and Reserve Affairs leads an organization of approximately 900 personnel who are responsible for managing manpower in the U. S. Marine Corps. Within M&RA, the Marine Corps' enlisted force planners have the important task of balancing requirements - billets - with resources – the Marines who fill them. While M&RA is the center of the manpower planning process for the Marine Corps, it is only one player out of several in this process. In calculating the number of required and forecasted reenlistments, coordination with other Marine Corps agencies is required.

The Marine Corps Combat Development Command (MCCDC) houses the Total Force Structure Division (TFSD) and Training and Education Command (TECOM). Each year TFSD determines the numbers of required personnel by MOS and grade, and their respective training requirements from designated representatives of each occupational field (occfield), called the occfield sponsor.  The Marine Corps currently has more than 30 occfields.  With TECOM's oversight and coordination with the many training pipelines for enlisted Marines, TFSD formulates the Marine Corps' Authorized Strength Report (ASR) using occfield sponsor inputs.  Marine Corps Recruiting Command (MCRC) and TECOM also provide inputs to the ASR concerning the accession of new recruits and their training pipelines.  The ASR summarizes the endstrength requirements for personnel by MOS and grade.  (Zamarripa, 2005)  An abbreviated depiction of the manpower planning process is shown in Figure 1.

Next, TFSD forwards the completed ASR to M&RA's Plans and Integration section (MPP-50).  In order to account for all Marines not currently serving in their primary MOS billets, analysts at MPP-50 forecast the numbers of Marines who have the status of patient, prisoner, trainee, and transient.  The quantities of Marines in these status categories are called "P2T2" estimates.  Analysts then subtract the appropriate quantities of P2T2 from each MOS and grade category in the ASR to give a realistic goal for planners to work toward.  The end product, after P2T2 adjustments to the ASR, is called the Grade Adjusted Recapitulation (GAR).

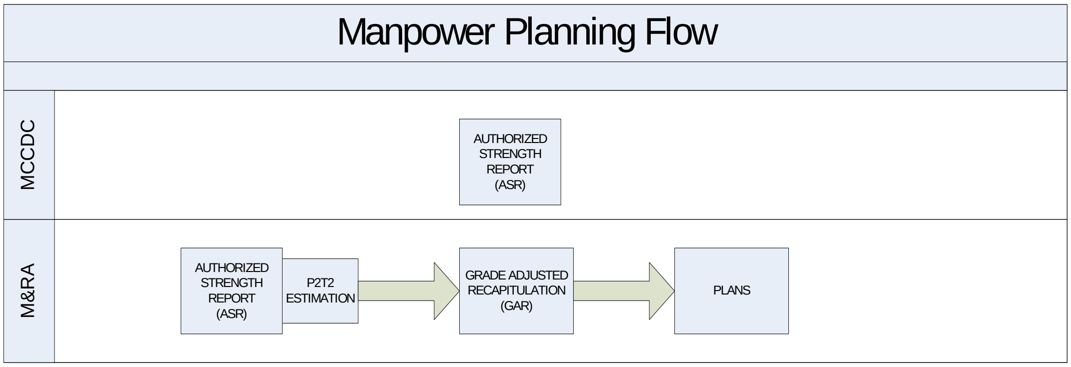
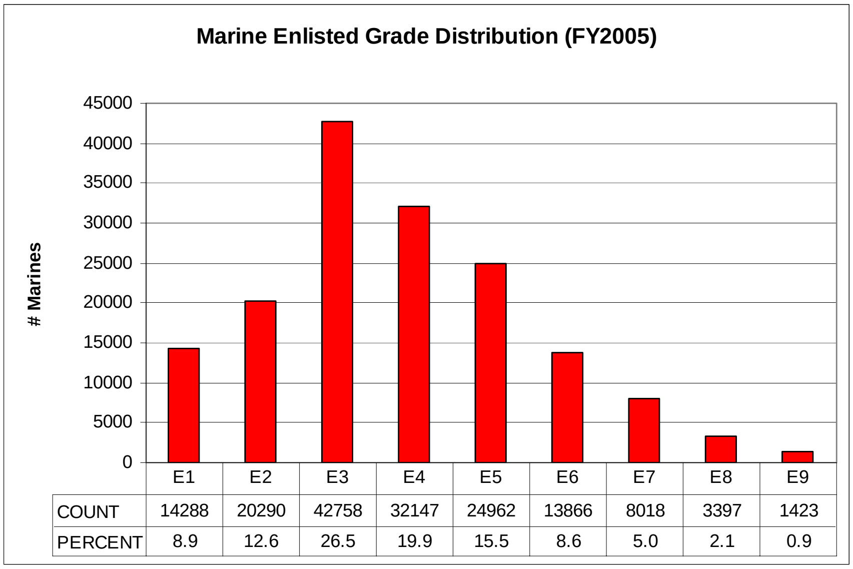

Manpower Planning Flow
MCCDC
AUTHORIZED STRENGTH REPORT (ASR)
M&RA
AUTHORIZED STRENGTH REPORT (ASR)
P2T2 ESTIMATION
GRADE ADJUSTED RECAPITULATION (GAR)
PLANS

Figure 1. Abbreviated manpower planning flow. (After: Zamarripa, 2005)

TFSD and M&RA work together to create a GAR for up to 5 years in the future, and enlisted force planners in the Enlisted Plans Section (MPP-20) use the GAR estimates three years into the future as a target force for the next fiscal year (FY).

### 2. The Marine Enlisted Force

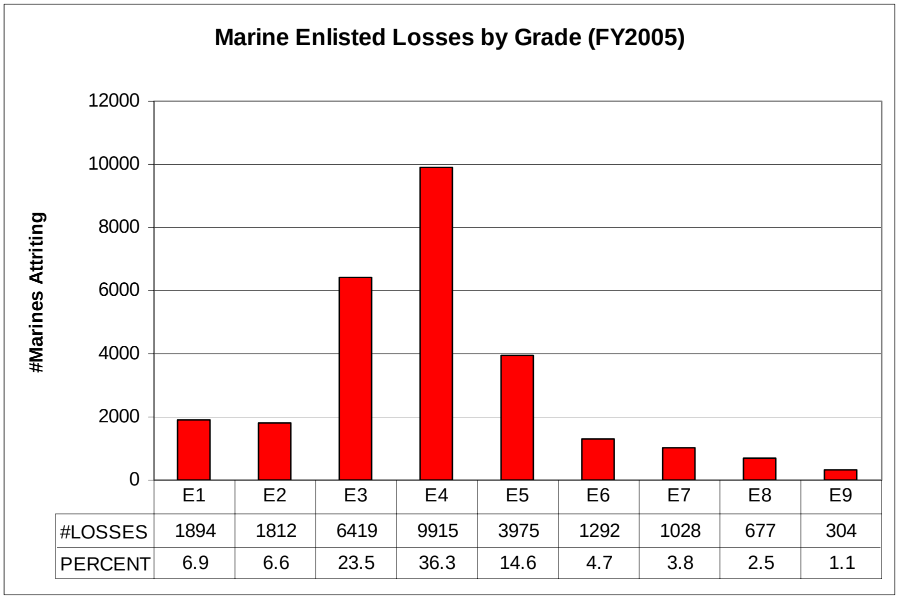
Roughly 30,000 new recruits enter the Marine Corps each year. Marines enter the enlisted force on contracts ranging from three to six years in length, with the most common lengths being three- and four-year contracts. The term “accessions” is used to describe the newly enlisted Marines. In order to maintain a force of roughly 161,000 enlisted Marines (see Figure 2 for the grade distribution for Fiscal Year 2005), separations from the Corps must be approximately equal to accessions. Obviously, this is not the case when the Marine Corps is attempting to increase or decrease its size.



| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 |
|---|---|---|---|---|---|---|---|---|---|
| COUNT | 14288 | 20290 | 42758 | 32147 | 24962 | 13866 | 8018 | 3397 | 1423 |
| PERCENT | 8.9 | 12.6 | 26.5 | 19.9 | 15.5 | 8.6 | 5.0 | 2.1 | 0.9 |

Figure 2. Enlisted Grade Distribution, Fiscal Year 2005.

Most of the Marine Corps’ personnel turnover takes place in the junior ranks, among those serving in their first enlistment contract. As shown in Figure 3, more than 70 percent of the Marine Corps’ enlisted personnel attrition occurs in the lowest four grades (E1, E2, E3, and E4).



| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 |
|---|---|---|---|---|---|---|---|---|---|
| #LOSSES | 1894 | 1812 | 6419 | 9915 | 3975 | 1292 | 1028 | 677 | 304 |
| PERCENT | 6.9 | 6.6 | 23.5 | 36.3 | 14.6 | 4.7 | 3.8 | 2.5 | 1.1 |

Figure 3. Enlisted Attrition by Grade, Fiscal Year 2005.

A Marine serving in his or her first enlistment (or contract) is said to be a "first-term" Marine. All other Marines, those who have served into a second enlistment or beyond, are called "subsequent-term" Marines for manpower planning purposes. The Marine Corps uses different policies regarding the separation of first- and subsequent-term Marines.

Most Marines nearing the end of their first term will not reenlist in the Corps. Those who wish to reenlist and serve in their particular MOS must be in an MOS with sufficient vacancies. Such vacancies are created by subsequent-term Marines that have been promoted or who have separated from the service. Manpower specialists call these vacancies for new second-term Marines "boat spaces." Of course, subsequent-term Marines also create vacancies for other subsequent-term Marines to fill. However, boat spaces are different from vacancies that exist for subsequent-term Marines.

The end of the first enlistment is the last point at which the Marine Corps can effectively separate a Marine without providing Involuntary Separation Pay to leave the service. This is true in all of the United States Armed Forces. If a Marine serves his or her second enlistment and attains at least six years of service, he or she is afforded Involuntary Separation Pay if he or she is forced to leave the service. (Marine Corps Order P1040.31J, 2004) Therefore, the point at which first-term enlistees must either depart the Marine Corps or reenlist is an area where planners focus a great deal of attention on vacancies (boat spaces in this case). It is extremely costly to order the separation of subsequent-term Marines, and also costly to re-train men and women into specialties other than their original ones. Therefore, the accurate calculation of boat spaces and first-term reenlistments required is critical.

A first-term Marine in good standing, and for whom there is no available boat space open in his original MOS, may apply for a lateral move to another MOS. Enlisted force planners create lateral move opportunities for qualified first-term Marines who wish to reenlist. Generally, this takes place in MOSs which are forecast to be undermanned in the coming fiscal year. In the case of lateral moves, each Marine must be re-trained into his or her new primary MOS by attending formal schooling. In most cases the Marine is promised a reenlistment bonus upon completion of the school and official receipt of the new MOS designation. Most lateral moves are executed by Marines at the start of their second enlistment. However, subsequent-term Marines are also sometimes permitted to make lateral moves.

In summary, the major differences between the first and subsequent-term components are:

- First-term Marines may be separated from the Corps without extra pay.
- Marines desiring reenlistment at the end of the first contract must have a specific vacancy to fill, that was created by the promotion or attrition of a second-term Marine.
- Second-term Marines, upon reaching 6 years of service, receive Involuntary Separation Pay when forced out of the Marine Corps for reasons other than substandard performance or criminal conduct.

### 3. Forecasts Required for Planning

Because of the differences just discussed, M&RA has used two separate models for determining the numbers of required reenlistments for first and subsequent-term inventories. The First Term Alignment Plan (FTAP) model was developed in 1991 by the Center for Naval Analyses. Its motivation grew from the need to reduce the overall size of the Marine Corps, while balancing the shrinkage of the junior and senior grades. In short, the FTAP model forecasts the promotion and reenlistment flows of the first-term population from one year to the next. The FTAP assumes that the target force (GAR) and flow rates remain unchanged from one year to the next. The FTAP model will be discussed in more detail in the next chapter of this thesis.

The model used for the subsequent-term Marine population is called the Subsequent Term Alignment Plan (STAP) model. It was developed at M&RA in 2001 as a tool to assist in planning for the career movements by those in their second term or beyond. The STAP model uses attrition rates in its population to forecast the next year's inventory of Marines before reenlistments occur. Forecasting these inventories enables enlisted force planners to distribute the Selective Reenlistment Bonus prudently in order to influence the retention of both first- and subsequent-term Marines in MOSs which are forecast to be undermanned.

In summary, the outputs of both the FTAP and STAP models are used to create a forecast, by grade and MOS, of the structure of the Marine enlisted force. Details such as the number of required of reenlistments can be taken from these models in order for planners to apply the appropriate influences (SRB and lateral moves) to appropriately shape the inventory of Marines.

The FTAP and STAP models were created roughly ten years apart. They utilize related, yet different methodologies. The FTAP model applies continuation rates by occupational field and years of service, and executes in a set of Excel spreadsheets. Using SAS, the STAP model applies attrition, retirement, and promotion rates to the inventories of Marines having grades E-5 (Sergeant) through E-7 (Staff Sergeant). Resident knowledge at M&RA enables planners to run these independent models twice a year to make forecasts and their resulting plans. However, the consensus at M&RA is that a new model would be beneficial for several reasons.

Ideally, a new model should consolidate the FTAP and STAP calculations into one, coherent source. Second, the new model should calculate the optimal distribution of the SRB budget each year. This part will be left for follow-on work. Third, the new model ought to be maintained and updated "in house" between the Enlisted Plans Section (MPP-20) and the Integration and Analysis Section (MPP-50), with no inputs required from outside agencies. M&RA established the title of Career Force Retention Model (CFRM) for the efforts leading to the new model.

# II. LITERATURE REVIEW

In November 2005, the United States Government Accountability Office (GAO) published a report entitled "DOD Needs Action Plan to Address Enlisted Personnel Recruitment and Retention Challenges." The report stated that "19 percent of DOD's 1,484 occupational specialties were consistently overfilled and 41 percent were consistently underfilled from FY 2000-2005." (Introduction, ¶2) Although it is very difficult to maintain occupational specialties at exactly the desired levels, the GAO's analysis indicating that certain specialties were consistently under- and overfilled suggests problems in the military manpower process. The problem can lie in several areas, some of which are the retention of qualified service members and misguided incentive programs to retain them. GAO also complained of a lack of useful information from the Armed Services about their incentive programs, which was not helpful in judging incentive effectiveness.

A well documented effort by North and Quester of CNA (1991) provides good background information on the scope and methodology used to create the FTAP model currently used. The methodology used prior to the FTAP model focused on the transition of first-term Marines into the career force by looking at transitions into the fourth through sixth years of service "band" (or interval). Changes needed to be made to this method based on contract lengths predominant at the time and the need to incorporate continuation rates throughout the entire span of the career force. Hence the FTAP model shifted to determining requirements in the fifth through twentieth years of service by occupational field as opposed to stopping at the sixth year of service.

"Managing the Enlisted Marine Corps in the 1990s Study: Final Report" (Quester and North, 1993) summarizes the work done by CNA from 1991 to 1993 in the Marine Corps manpower field. One purpose of the study was to gain insight into the reenlistment decision at the Marines' end of contract. CNA created a longitudinal data file for all Marines from 1980 to 1991 in the sixth through fourteenth years of service, containing demographic information such as race, gender, and marital status, and SRB multiple offered. Also included were variables describing a Marine's service such as an indicator of contract extension, grade, and years of service. CNA included numerical economic variables such as military to civilian pay ratio and national unemployment rates as applicable to the age groups of Marines in their data set.  Although it was not explicitly stated, the reader is left to assume that CNA used logistic regression in this study given the content of their other, similar studies.

The Marine Corps uses time in service to assign personnel to an SRB Zone. There are three zones as defined in Table 2 below.

Table 2. SRB Zones, determined by time in service.

| ZONE | FROM | TO |
|---|---|---|
| A | 17 MONTHS | 6 YEARS |
| B | OVER 6 YEARS | 10 YEARS |
| C | OVER 10 YEARS | 14 YEARS |

CNA found that in both Zones B and C, married Marines were respectively 11 percent and four percent more likely to reenlist than unmarried Marines.  Further, Hispanic and African-American Marines were more likely to reenlist than those with other racial backgrounds.  Further results showed that raising the SRB payment by a multiple of one increased enlistment rates by about seven percent and five percent, respectively for Marines in SRB Zones B and C.

CNA also published “Cost Benefit Analyses of Lump Sum Zone A, Zone B, and Zone C Reenlistment Study: Final Report.” (Hattiangadi et al., 2004) Using a longitudinal data set similar to the one cited in previous work, logistic regression was used to determine reenlistment propensities by occupational field and reenlistment zone. The data set for this study included all Marines facing reenlistment decisions between 1985 and 2003.  Similar demographic variables were used with the addition of occupational field, AFQT score, and whether the reenlistment occurred between 1992 and 1997 (a force reduction period).  Noted in the study was that it marked the first such effort since the implementation of the lump sum bonus in 2000, instead of the former system of installment payments of SRBs.

This work hypothesized that the effects of race and family status would be useful in forming their models. Furthermore, the assumption about AFQT scores and the sensitivity to monetary incentives (SRB) were presumed:

> Research indicates that ability, as measured by the AFQT score, has a large effect on reenlistment rates ... Servicemembers' sensitivity to compensation increases can vary with AFQT score. Specifically, Marines with higher AFQT scores are less likely to reenlist but may be more sensitive to SRBs ... we interact the SRB bonus level with AFQT to see if those with AFQT scores in the top half ... react differently to positive SRB offers. (p. 44)

Interesting results of this study showed that SRB multiple was a significant factor in the logistic regression models, and that its marginal impact on reenlistment rates was highest in Zone B (Marines in YOS six to fourteen) with a gain of 7.2 percent per SRB multiple increase. The SRB effect was slightly less in Zone A (6.6%) and Zone C (3.5%). Racial variables "Black" and "Hispanic" had statistical significance at the 99 percent level, showing increased enlistment likelihood of varying degree between reenlistment zones, for Marines belonging to these groups. Gender showed no effect on reenlistment in Zone A, and small marginal effects in Zones B and C.

A 1999 Naval Postgraduate School (NPS) thesis, by Australian Army Major Karl S. Delany, used logistic regression for determining factors important in reenlistment to a specific cohort (determined by AFQT score > 50 and contract length of three or four years) of United States Army soldiers in their first-term between 1992 and 1996. Delany used many of the same predictor variables that were used in CNA's 2004 study. He did not incorporate economic indicators in his model, but did use measurements of age, education, education incentive (Army College fund, or ACF), and whether the soldier was in a technical field.

Results from Delany's research suggested that length of initial contract, pay grade, family status, race, and AFQT score were the most significant predictors in his logistic regression model. Delany's results unexpectedly indicate that receiving a reenlistment bonus caused a two-percent reduction in the probability of reenlistment. No validation of the model was conducted, as it was used for determining significant factors in the reenlistment decision, not as a predictive tool for individual reenlistment.

A 1997 Naval Postgraduate School thesis by U.S. Navy Lieutenant Commander Terrence S. Purcell used Classification and Regression Trees (CART) to predict the category of attrition of soldiers in the U. S. Army. This research explored the use of CART as a legitimate tool for data exploration and prediction. Purcell used a subset of Army soldiers in the serving in any of the years 1983 to 1988 to create classification trees in S-Plus data analysis software. The trees were grown without restriction in size to reveal structure and relationships within the data. Next the trees were cross-validated and pruned based on the cross-validation diagnostic information, to prevent overfitting the data used to create the models. Terminal nodes of the classification trees indicated the numbers of soldiers classified and the proportions of each classification type within the node. Three types of attrition were classified, along with "Not" lost by the end of the first term, indicating that the soldier reenlisted for a second contract.

Only categorical explanatory variables were used in Purcell's research. The information contained in these variables was similar to that in works cited above, including the following variables: length of service term, AFQT, education background, gender, and race. For use in the tree models, AFQT scores were used to create four categorical variables based on percentile of score for each individual.

The variable partitioning performed by the tree models offered good insight into what factors might determine the nature of soldiers' separation from the service. Purcell suggests in that using "attributes [categorical variables] with few levels results in terminal nodes with very broad characteristics. By increasing the levels of a particular attribute, the terminal nodes will be more tightly defined." (p. 59) He further clarifies that the purpose for making a tree model (i.e., prediction or data exploration) should determine the proper extent of pruning the trees. With several models created using CART, variables which consistently contributed the most in correctly predicting the type of soldier attrition were race, length of enlistment contract, and gender. In some cases, other variables such as AFQT score and education level contributed to the trees' predictive ability, depending on the extent of pruning, and other variables included in the model.

Another Naval Postgraduate School thesis, by U.S. Navy Lieutenant William B. Hinson (2005), used classification trees and logistic regression to predict students' success following foreign language training at the Defense Language Institute. In evaluating the set of predictor variables in the data, Hinson used a classification tree to help determine which variables were important in prediction.

Hinson also used information from the tree’s binary splits to make modifications to variables which were useful in developing a logistic regression model. One modification made was the collapsing of a categorical variable with five levels into three. This was done because two of the levels of the original variable applied to a very small proportion of the data.

THIS PAGE INTENTIONALLY LEFT BLANK

# III. METHODOLOGY

This section reviews the general concepts behind Classification and Regression Trees (CART) and logistic regression, the two primary techniques used for analysis in this thesis. CART is used to assist in variable selection for the predictive model. Logistic regression is then used to predict the number of Marine reenlistments by MOS and grade. Model goodness-of-fit is described in the last part of this chapter.

## A. OVERVIEW OF CART

CART is a useful, non-parametric, tool for data exploration and predictions in classification and regression. This description of CART follows Hand, Mannila, and Smyth (2001), who summarize three basic attributes of the CART algorithm as: "(1) a tree model structure, (2) a cross-validated score function, and (3) a two-phase greedy search over tree structures ('growing' and 'pruning')." (p.151) This thesis focuses on the use of classification trees.

Typical output from software having a CART method (or another, similar algorithm) includes a diagram of a tree structure. At the top of the tree is the root node, which theoretically contains all observations, and hence classifications of the data. Below the root node, a hierarchy of nodes is displayed, which represents binary split decisions based on recursive partitioning of variables. These nodes also "contain" observations from the data set, and have the attributes described by the tree's path (a data vector) ending at that given node. At each node, the algorithm determines two important things: (1) which variable to split on, and (2) at what threshold.

The variables in the input data set may be categorical, real, or integer-valued. The threshold for each binary split is determined by the goal of minimizing a loss function. The loss function used in this research is deviance.

Figure 4 is an example of a pruned classification tree resulting from running the CART algorithm in S-plus to predict reenlistment for subsequent-term Marines in FY2001. In Figure 4, variable splits are indicated on arcs. In the example, a value of '1' below a rectangular, terminal node indicates that Marines having attributes that follow that path are predicted to reenlist. A ‘0’ below a terminal node indicates a group of Marines predicted to leave the service.

Misclassification proportions are given in the fraction below each rectangular, terminal node. The usual loss function for splitting is deviance, which is a log-likelihood function. The tree is grown using deviance as a measure of impurity in each node, and as an overall score for the model. This is related to, but not the same as, using misclassification rate when pruning the tree.

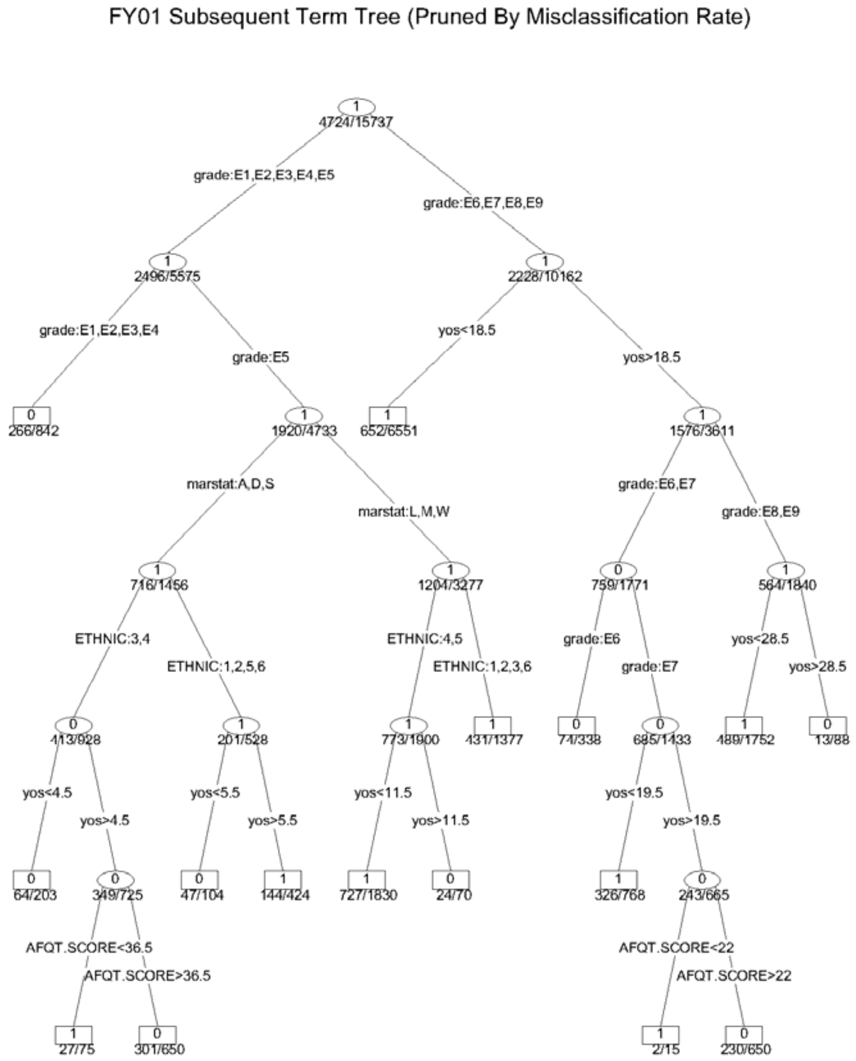

FY01 Subsequent Term Tree (Pruned By Misclassification Rate)

Figure 4. Example of a Pruned Classification Tree.

Once a relatively large classification tree is grown to fit the data, we cross-validate it and prune it to ensure its generality and thus, the ability to predict observations from data not used in making the tree. We will cross-validate and prune based on achieving a minimal misclassification rate paired with its associated, reasonable tree size. In CART, cross-validation is a means to ensure that a tree is grown that can predict reasonably well using new data that was not used in growing the tree. "Prune" means to reduce the size of the tree by removing nodes that contribute the least in predicting. Hand, et al. define the misclassification loss function as

$$\sum_{i=1}^{n} C\left( y(i), \hat{y}(i) \right)$$

where *y(i)* is the actual class for the *i*th data vector, and $\hat{y}(i)$ is the predicted class (p. 147). When $y(i) \neq \hat{y}(i)$, the cross-validation algorithm counts a loss of one.

A tree can be grown to have as many nodes as necessary to correctly classify each observation in the data. Such a large tree can be difficult to interpret. This overgrown tree might be useful in understanding structure of the variables in a data set. (Purcell, 1997) However, overgrown trees rarely have predictive ability. Because an overgrown tree perfectly fits the data from which it was grown, it will not often correctly classify data from other data sets with great success. This is where cross-validation comes in. Hand, et al. state that cross-validation "allows CART to estimate the performance of any tree model on data not used in the construction of the tree – i.e., it provides an estimate of generalization of performance." (p. 149)

In tree cross-validation, the data is equally split into *N* subsets. Because it is a reasonable default in S-Plus, *N*=10 subsets were used in this research. Tree models are built iteratively using *N* minus one (all but one, or nine) of the subsets, and the misclassification rate is determined by the model's prediction on the tenth, or "left-out" data set. Tree models of different sizes are created, and then scored based on misclassification rate. Using software such as S-Plus or Clementine, one can determine an ideal tree size based on the best size and misclassification pairing. Once a best size of the tree (measured by the number of nodes) is determined, the overgrown tree is pruned to that size.

The "`tree`" method in S-Plus was used in this research. It is a recursive partitioning algorithm that implements the CART method just described.

## B. OVERVIEW OF LOGISTIC REGRESSION

Logistic regression is a widely-used statistical methodology that is particularly useful for estimating the probability of a binary (dichotomous) event given other information. In simple linear regression, we can form a relationship between a set of predictor variables and a quantitative response variable. Devore (2004) defines the usual notation for doing this as the model equation, given by

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

(p. 500). Here Y is the response variable, $\beta_i$ is the slope parameter (sometimes called the coefficient), and $\varepsilon$ is an error term. The generalization of simple linear regression is multiple regression which has multiple predictor variables (*x*s) and slope parameters. Coefficients of the linear regression model are found by minimizing the residual sums of squares. The reader is referred to Devore for a more in-depth discussion of this model.

The equation above is sufficient for modeling data for which the real-value response interval lies in $(-\infty, \infty)$. A response variable that is dichotomous is usually coded as a 0 or 1 in the data. The linear regression model is inappropriate in this case because it would most likely lead to predictions outside of the interval [0,1]. Further, linear regression maintains the requirement of constant variance in the residuals. This residual variance structure cannot be maintained when using a dichotomous response variable.

Logistic regression provides a solution to these problems. This description of logistic regression parameters and notation follows Fleiss, Levin, and Paik (2003). The probability of an event occurring (reenlistment in this case) is called *P*. We define the log odds (often called the logit) transformation of *P* as

$$\text{logit}(P) = \ln\left(\frac{P}{1-P}\right) \text{ (Fleiss, et al., p. 284)}$$

and logistic regression then models the logit as a linear function of the predictor variables

$$\text{logit}(P) = \beta_0 + \beta_1 x + \varepsilon.$$

The logit has no restrictions on its value (i.e., it can lie anywhere on the interval $-\infty, \infty)$. Furthermore, for a given value of *x*, if we calculate $\lambda = \beta_0 + \beta_1 x + \varepsilon$ then, again for that value of *x*, we can estimate the probability of reenlistment as

$$P = \frac{e^{\lambda}}{1+e^{\lambda}} = \frac{1}{1+e^{-\lambda}} \text{ (Fleiss, et al., p.284).}$$

As with linear regression, logistic regression can also be generalized to have multiple predictors.

In this thesis, *P* is the probability of a Marine reenlisting at the end of an enlistment contract. The attributes of the Marine, such as demographic information and service characteristics, are accounted for in the data vector that represents the individual. Finally, the probabilities for each Marine's reenlistment are summed for each grouping of MOS and grade to predict the number of reenlistments in each particular MOS and grade combination.

## C. ASSESSING MODEL GOODNESS OF FIT

Several different techniques may be used in logistic regression to assess the usefulness of a model and the choices made in selecting variables. In the reviewed literature, one of the usual methods for building a logistic regression model is to evaluate the statistical significance of each of the predictor variables. This is measured using a chi-square statistic, and the *p*-value for the resulting statistic given. Predictors meeting the pre-determined, required level of statistical significance are chosen to remain in the model.

Such an approach is based on the idea of sampling, in which a sample of a population is obtained and the statistical significance of a variable means that it is useful in inferring some characteristic or relationship from the sample back to the entire population. The work in this thesis differs in two important respects from this scenario. First, the models are built using the entire population's data for a given year. Second, the goal is to use the model from one year to *predict* the next year.

That is, the goal of this thesis is to accurately predict the number of reenlistments by MOS and grade. Hence, in this research we are not interested assessing the model using *p*-values and other traditional methods. Rather, we are interested in simply assessing how well the model predicts. And, given that we have sufficient historical data from which we can create models and then make predictions for years for which we already know the outcome, the relevant measure of fit is to compare the predictions to the actual data – the closer the prediction the better.

To this end, we use a chi-square-like statistic to measure the overall fit of the logistic regression model's predictions to the data. As was just described, for each MOS and grade, the actual number of reenlistments, or ground truth, is known for any given year past. To measure the predicted deviations from ground truth, the squared difference is calculated between the number of predicted and actual reenlistments, and divided by the predicted number of reenlistments. This will be called "AVG DIFF," for the average squared difference in the model's output. This calculation is made for each cell of Marines, indexed by MOS and grade. For a measure of how well, overall, a model fits the many MOS and GRADE cells of Marines, we sum all of the AVG DIFF measurements. This statistic will be referred to as AVGDIFF$_{\text{MODEL}}$. In short, the calculation we use for assessing overall fit of the model is defined by:

$$AVGDIFF_{MODEL} = \sum_{MOSGRADE} \frac{\left(predicted\,\#\,reenlistments - actual\,\#\,reenlistments\right)^2}{predicted\,\#\,reenlistments}\,.$$

The model's predictor variables were then selected based on: (1) insight gained from classification trees and literature and (2) satisfying the goal of minimizing AVGDIFF$_{\text{MODEL}}$. All data manipulation and regression work was done using SAS software. Examples of calculations from the model's output are shown in the results section of the last chapter.

# IV. DATA SET AND VARIABLES

## A. DESCRIPTION OF THE DATA SET

The data set used for predicting reenlistment contains all enlisted Marines from the years 1998 to 2005. The Marine Corps' Total Force Data Warehouse (TFDW) provided the database from which to draw the data set. End-of-month snapshots, called "sequences," of the entire Marine Corps population are stored in TFDW. This data exists for all years from 1988 until the present. For this thesis, the years of interest include 2001 to 2005 for purposes of developing a prediction model for number of reenlistments in the most recent year, 2005.

Using SAS we imported all of the snapshots from TFDW and merged them into one longitudinal data set. The longitudinal data set contains one row, or observation, for each Marine who was in the service at any point between 1988 and 2005. Each observation for a Marine is taken at the end of the fiscal year (30 September of each year). The data set appears sparse, since it contains missing values for each Marine who was not in the service during a particular year.

The TFDW longitudinal data set provided our demographic predictors, both fixed and time-varying, such as race and number of dependents. It also contains time-varying service-related variables for each Marine, such as MOS and years of service (YOS). Missing values were imputed by going back one year at a time for three consecutive years and using the most recent data to fill in for data missing in the current year. For example, if a Marine had a missing value for MOS in 2004, the value from 2003 filled the gap, given that it is not missing. If the value for MOS in 2003 was also missing, then the value from 2002 was used, and so on, reaching back to 2001. Generally speaking, going back one year filled in the majority of the missing data and all missing data was corrected by going back no more than three years.

Two other data sets were examined for the purposes of gaining more information about Marines prior to a reenlistment decision. Deployment data for Marines in TFDW was available. Unfortunately, the deployment data set contained deployment information only for Marines who were still in the service at the end of FY2005. Therefore, not enough data was available to use deployment as a predictor, since information was required, but not present, for those who left the service. In determining SRB eligibility, data was merged from the administrative messages posted on the Marine Corps' website (www.usmc.mil). During the last part of each fiscal year, the SRB message is released, stating which MOSs and eligibility zones would receive specified bonuses upon reenlistment. This data was imported into SAS and merged with the demographic data set. Since a Marine's time in service can be calculated from the data in TFDW, we were able to closely approximate the SRB eligibility zone for each Marine, and merge the appropriate bonus offer for that year when applicable. This will be discussed more in the next section.

Table 3 shows an example of a Marine who served from 2001 through 2005 in the longitudinal data set (LDS). Notice that variables such as "SEX" and "ETHNIC" (the race and ethnic code) do not change with time. However, variables such as MOS, grade, YOS, and marital status may change from one year to the next. These variables were indexed by year in the LDS. Due to the large number of variables, not all are shown in Table 3. In the first observation of the example below, the Marine's YOS variable was incremented four times over the four year-period, and his MOS and marital status changed. Longitudinal data indexed by the years 2002 through 2004 are not shown because of space constraints.

Table 3. Examples of Marines serving from 2001 through 2005 in LDS.

| OBS | SEX | ETHNIC | AFQT | MOS2001 | GRADE2001 | YOS2001 | MAR2001 | … | MOS2005 | GRADE2005 | YOS2005 | MAR2005 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | M | 1 | 90 | 0311 | E3 | 3 | S | … | 0369 | E6 | 7 | M |
| 2 | M | 4 | 78 | 3051 | E6 | 6 | M | ... | 3051 | E6 | 10 | M |
| 3 | F | 2 | 88 | 6113 | E5 | 5 | M | ... | 6113 | E6 | 9 | M |

In building classification trees and logistic regression models, only observations from the years of interest were used. In building the logistic regression model for predicting 2005 reenlistments, only two years' worth of information, per Marine, were required and extracted from the LDS. Variables that can change over time, such as

GRADE, were taken from the most recent end-of-year snapshot prior to the Marine's reenlistment year. If the Marine was eligible for reenlistment in FY2005, variables from the end of FY2004 were used to predict the reenlistment probability in logistic regression. Using the end of the FY prior to the reenlistment year makes intuitive sense, and was the best option due to the end-of-year "snapshot" composition of the LDS. The end of the FY prior to the reenlistment decision is virtually the beginning of the next year, which is the reenlistment year.

## B. INTRODUCTION TO DATA SET VARIABLES

### 1. Dependent Variable

Our logistic regression model was constructed in order to predict the reenlistment of a Marine during FY2005, given that he or she reached the end of his or her enlistment contract during that year. With this in mind we first determined whether the Marine had an End of Current Contract (ECC) date within FY2005. If a Marine's ECC date fell between October 1, 2004 and September 30, 2005, then the Marine was classified as being "eligible" for reenlistment during FY2005. The term "eligible" is solely based upon the ECC date of the Marine, not on whether that Marine is qualified to reenlist, or even recommended for reenlistment by his or her chain of command (which is a requirement to reenlist). Counting the number of previous contracts completed is required for each Marine in order to determine the Marine's status as a first- or subsequent-term population member.

Once a Marine was classified as eligible for reenlistment, the next task was to determine whether or not he reenlisted. If the Marine was not present at the end of the fiscal year where reenlistment eligibility took place (i.e., at the end of FY05 in this case), then a binary variable, called ELIGREEN1, was coded as 0 for "did not reenlist." If he or she was present at the end of the fiscal year, the variable was coded as 1, for "reenlisted."

However, note that because enlisted Marines may request extensions to their contract, it is possible that a Marine could be classified as having reenlisted, when he actually was in an extension of his most recent contract. No data was available that explicitly indicated a Marine serving an extension of a contract; therefore, Marines who reenlisted and extended were combined with respect to our predictions.

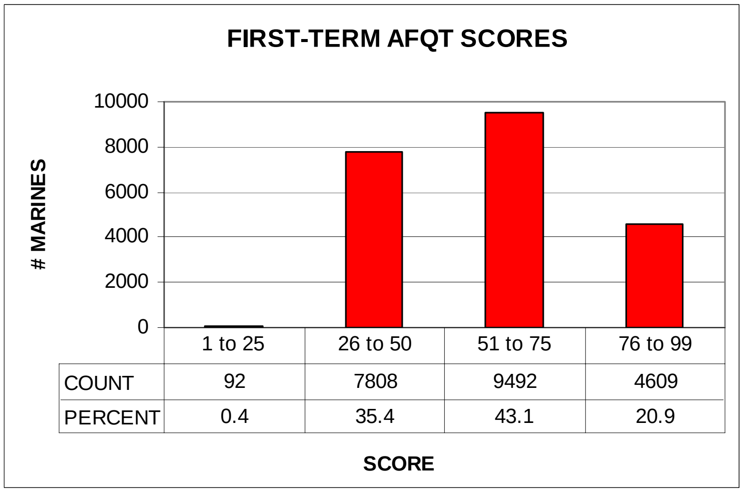
Indeed, what we are actually predicting is whether a Marine *continued* on active duty in the next year and hence we are using the term “reenlist” very loosely – “retained” would be a better descriptor. This is very much a mixture of “apples” and “oranges,” but as we just described, a mixture that we could not separate with the data that was available. From a modeling perspective, such a separation might be very useful for building a more accurate model, but from the practical perspective of M&RA, differentiating between the two groups is not material since M&RA simply needs to know the number of reenlistment eligible Marines that will be around in the following year (filling spaces).

### 2. Demographic Variables

The demographic variables are discussed next. Bar charts provide information about the group of Marines who were eligible for reenlistment during FY2005. In some cases, separate charts are shown for first-term and subsequent-term populations of Marines. All bar charts shown describe only Marines who were eligible for reenlistment during FY2005, unless otherwise stated. Note that the vertical ($y$) axes of the bar charts have different scales, as the first-term reenlistment-eligible population is larger than that of the subsequent-term population. This is always the case. Data for the charts are from a TFDW query for September 30, 2004 (Sequence Number 121) for both first- and subsequent-term Marines.

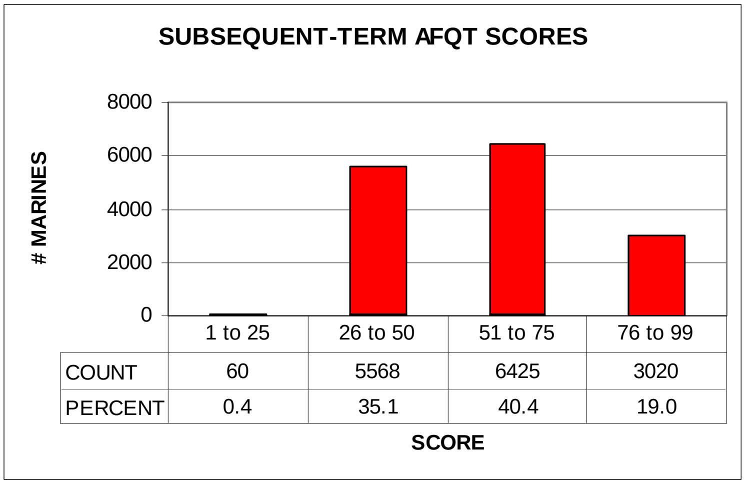
#### *a. AFQT_SCORE*

This variable represents the Marine’s score on the Armed Forces Qualification Test, a standardized test used by all military services to forecast an individual’s likely adaptation to military training and instruction. AFQT scores range from 1 to 99. The AFQT_SCORE variable had less than one-percent missing values in the first-term population, and roughly five-percent missing for the subsequent-term population. Despite missing values, the variable was still examined to determine if it improved model accuracy (it did not). Figures 5 and 6 show the distributions of AFQT scores.



| | 1 to 25 | 26 to 50 | 51 to 75 | 76 to 99 |
|---|---|---|---|---|
| COUNT | 92 | 7808 | 9492 | 4609 |
| PERCENT | 0.4 | 35.4 | 43.1 | 20.9 |

Figure 5. First-term population AFQT scores, end FY2004.

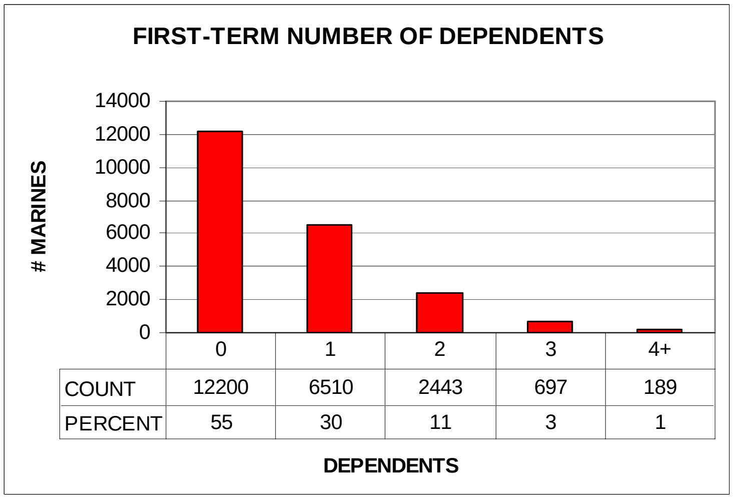


| | 1 to 25 | 26 to 50 | 51 to 75 | 76 to 99 |
|---|---|---|---|---|
| COUNT | 60 | 5568 | 6425 | 3020 |
| PERCENT | 0.4 | 35.1 | 40.4 | 19.0 |

Figure 6. Subsequent-term population AFQT scores, end FY2004.

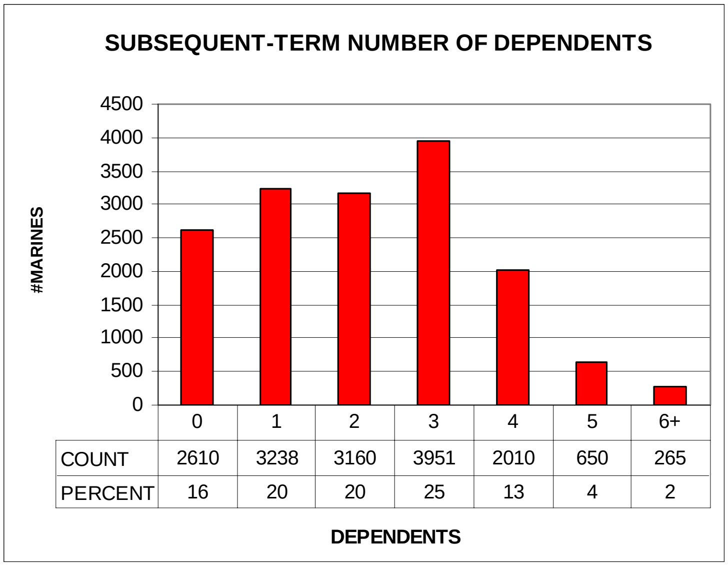
### *b. DEPSTAT*

"DEPSTAT" represents the number of dependents of the Marine. Figures 7 and 8 show the distribution of DEPSTAT for first-term and subsequent-term enlisted Marines at the end of FY2004 – the same data used for predicting reenlistments during FY2005.



| | 0 | 1 | 2 | 3 | 4+ |
|---|---|---|---|---|---|
| COUNT | 12200 | 6510 | 2443 | 697 | 189 |
| PERCENT | 55 | 30 | 11 | 3 | 1 |

Figure 7. Number of dependents for first-term Marines, end FY2004.



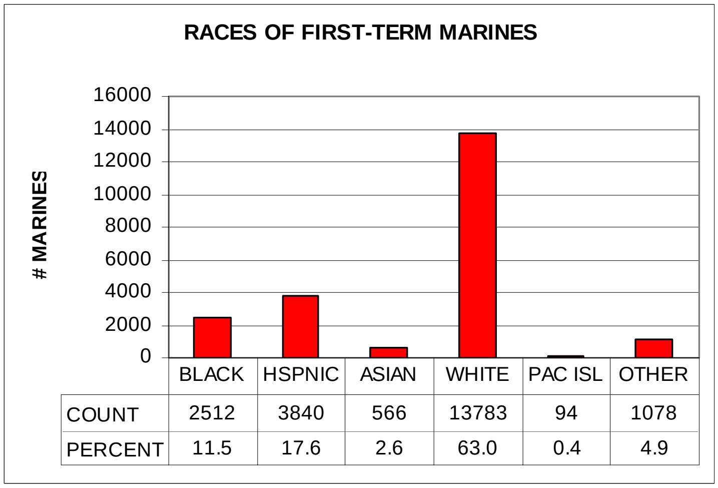
| | 0 | 1 | 2 | 3 | 4 | 5 | 6+ |
|---|---|---|---|---|---|---|---|
| COUNT | 2610 | 3238 | 3160 | 3951 | 2010 | 650 | 265 |
| PERCENT | 16 | 20 | 20 | 25 | 13 | 4 | 2 |

Figure 8. Number of dependents for subsequent-term Marines, end FY2004.

In the data set containing subsequent-term Marines, those extending their original enlistment contracts probably account for as many as 1,000 (8 percent) of the observations. These special cases are most commonly found in the E3 and E4 grades. Since we had no data clearly indicating that a Marine is an "extender," these were left in the subsequent-term population for modeling. Extenders fall in a gray area between first and subsequent term for modeling purposes.

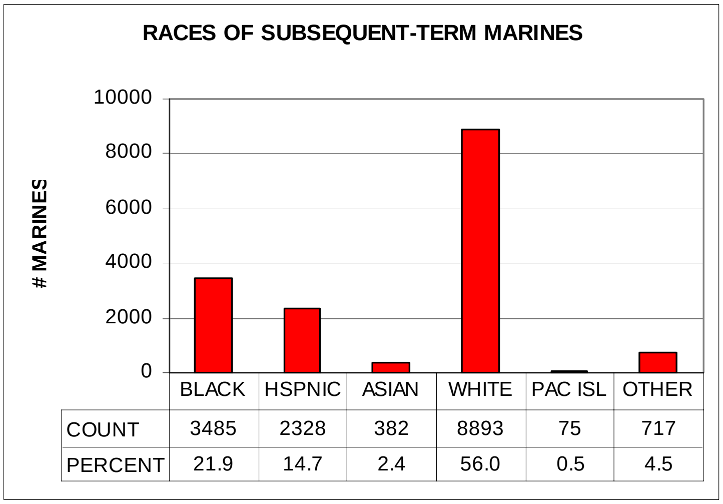
### *c. ETHNIC*

Race and ethnicity are coded separately in TFDW and, taken together, comprise dozens of possible combinations. For analytical purposes, these were collapsed into six racial/ethnic groupings. Collapsing the race and ethnic codes allowed the classification of many Marines who were classified as “Other” by their Race Code. For example, Hispanic Marines were generally classified as “White” in the Race Code. Since Hispanic Marines make up a substantial proportion of the Corps, it was useful to ensure that they were represented properly in a variable. On the next page, Figures 9 and 10 show the distribution of Marines for the variable ETHNIC.



| | BLACK | HSPNIC | ASIAN | WHITE | PAC ISL | OTHER |
|---|---|---|---|---|---|---|
| COUNT | 2512 | 3840 | 566 | 13783 | 94 | 1078 |
| PERCENT | 11.5 | 17.6 | 2.6 | 63.0 | 0.4 | 4.9 |

Figure 9. Racial composition of first-term Marine population, end FY2004.

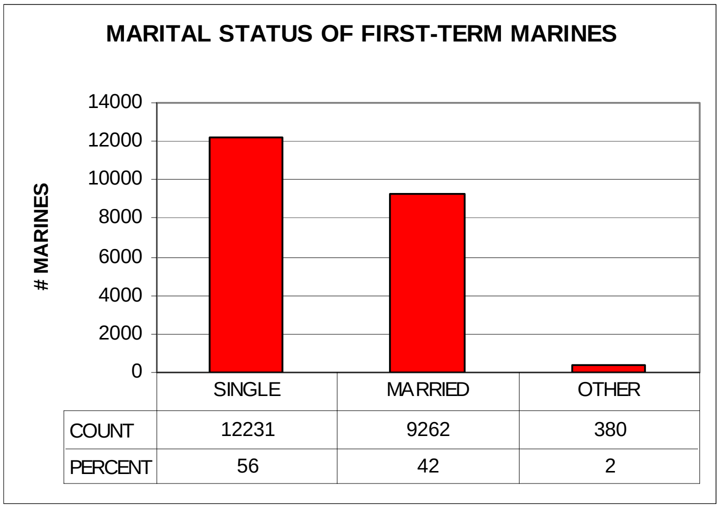


| | BLACK | HSPNIC | ASIAN | WHITE | PAC ISL | OTHER |
|---|---|---|---|---|---|---|
| COUNT | 3485 | 2328 | 382 | 8893 | 75 | 717 |
| PERCENT | 21.9 | 14.7 | 2.4 | 56.0 | 0.5 | 4.5 |

Figure 10. Racial composition of subsequent-term Marine population, end FY2004.

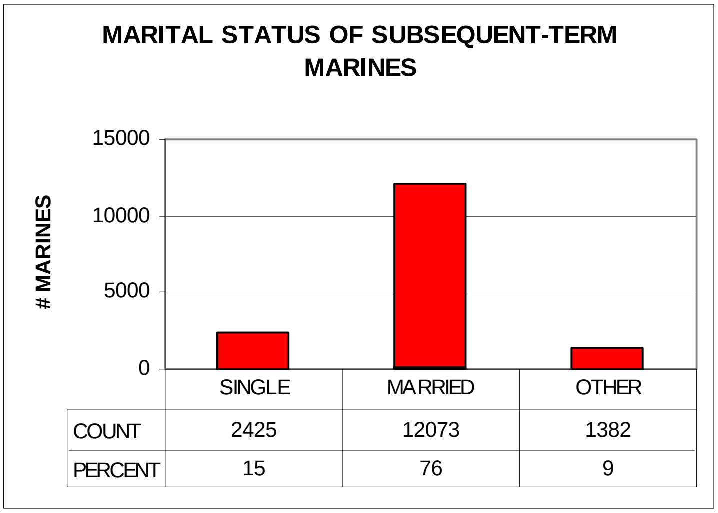
### *d. MARSTAT*

"MARSTAT" is the marital status of the Marine on the last day of the fiscal year prior to the end of current contract year. Levels of this variable include married, single, legally separated, divorced, annulled, and widowed. Figure 11 and Figure 12 show Marines' marital status from the LDS for FY05. All categories except "Married" and "Single" were grouped under "Other" due to their low numbers.



| | SINGLE | MARRIED | OTHER |
|---|---|---|---|
| COUNT | 12231 | 9262 | 380 |
| PERCENT | 56 | 42 | 2 |

Figure 11. Marital status of first-term Marines, end FY2004.

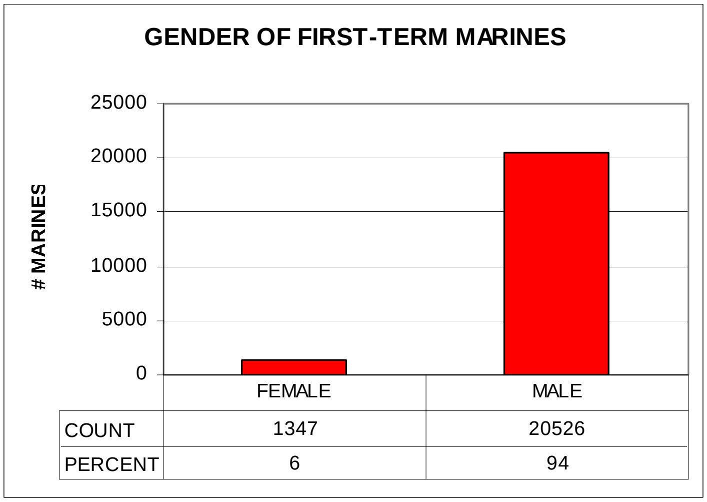


| | SINGLE | MARRIED | OTHER |
|---|---|---|---|
| COUNT | 2425 | 12073 | 1382 |
| PERCENT | 15 | 76 | 9 |

Figure 12. Marital status of subsequent-term Marines, end FY2004.

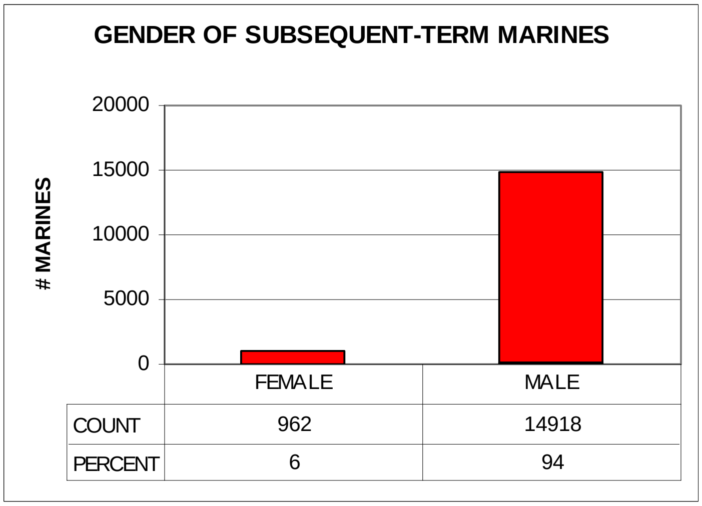
### *e. SEX*

This variable represents the gender of the reenlistment-eligible Marine. Figures 13 and 14 show the distribution of males and females in the first- and subsequent-term populations.



| | FEMALE | MALE |
|---|---|---|
| COUNT | 1347 | 20526 |
| PERCENT | 6 | 94 |

Figure 13. Gender of first-term Marines, end FY2004.



| | FEMALE | MALE |
|---|---|---|
| COUNT | 962 | 14918 |
| PERCENT | 6 | 94 |

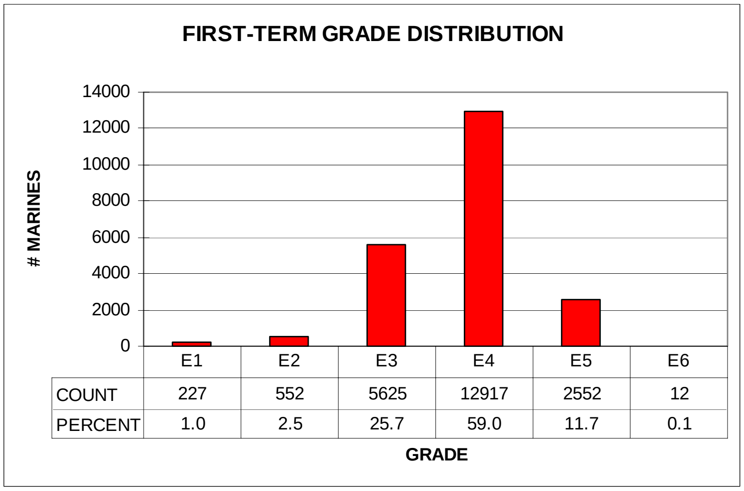
Figure 14. Gender of subsequent-term Marines, end FY2004.

## 3. Service-Related Variables

This section summarizes the six service-related variables in the LDS. Each of these variables appears once for each year in the Marine's row of the LDS.

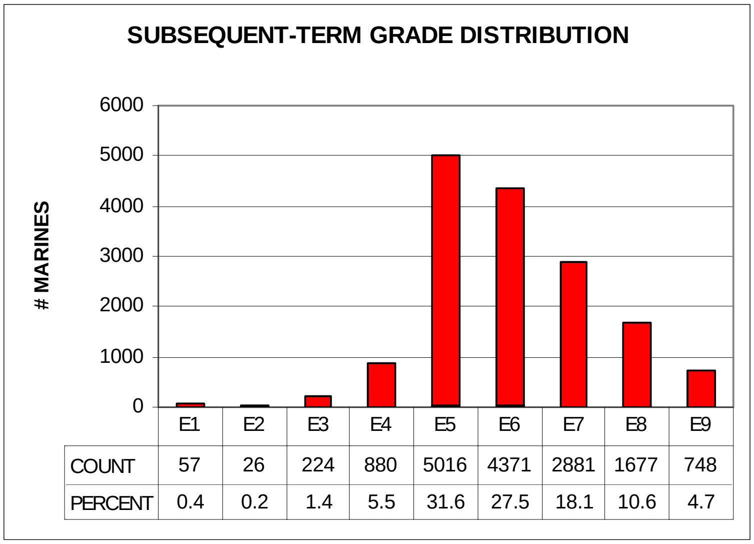
### *a. GRADE*

As in the other Armed Services, Marine Corps enlisted pay grades start at E1 and end at E9. Figure 15 and Figure 16 show the distribution of Marines eligible for reenlistment in 2005, by grade. For the first-term population, only Marines having between two and six YOS (inclusive) were included. Furthermore, only Marines having GRADE between E1 and E6 were included. These grade and years of service criteria led to the omission of 154 observations--a loss of less than one percent from the data set.



| | E1 | E2 | E3 | E4 | E5 | E6 |
|---|---|---|---|---|---|---|
| COUNT | 227 | 552 | 5625 | 12917 | 2552 | 12 |
| PERCENT | 1.0 | 2.5 | 25.7 | 59.0 | 11.7 | 0.1 |

Figure 15. Grade distribution of first-term Marines, end FY2004.

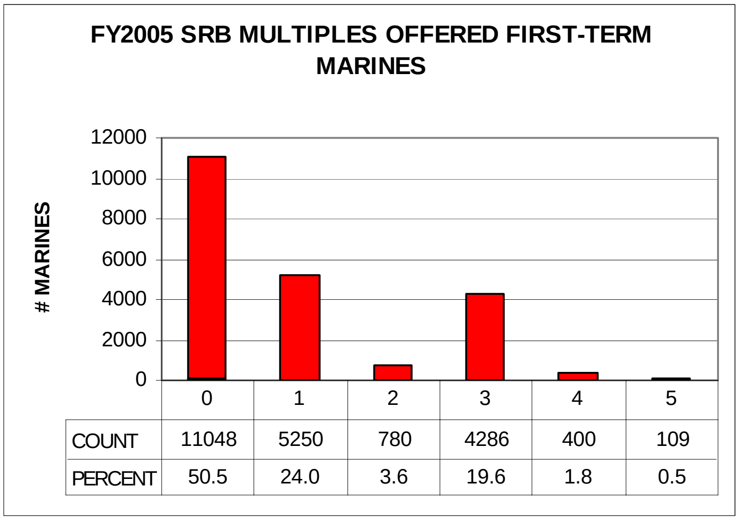


| | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 |
|---|---|---|---|---|---|---|---|---|---|
| COUNT | 57 | 26 | 224 | 880 | 5016 | 4371 | 2881 | 1677 | 748 |
| PERCENT | 0.4 | 0.2 | 1.4 | 5.5 | 31.6 | 27.5 | 18.1 | 10.6 | 4.7 |

Figure 16. Grade distribution of subsequent-term Marines, end FY2004.

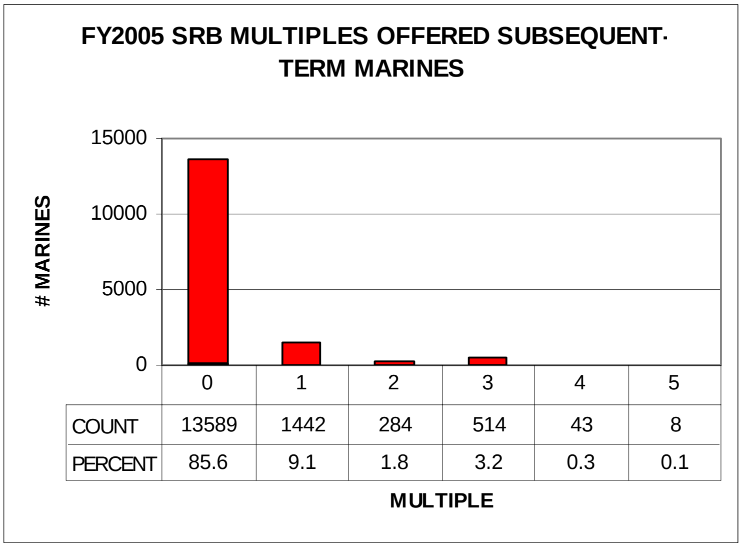
### *b. SRBELIG*

SRBELIG shows the Marine's SRB multiple which was offered his or her SRB Zone and MOS, as defined in Chapter II, Table 2. The Marine's SRB Zone, as stated earlier, was used to merge the data from historical SRB offerings. For example, if Marines in MOS 0311 and SRB Zone A, were offered a bonus multiple of three for reenlistment, then SRBELIG was set to 3 in the LDS for that particular year. Not all Marines belonging to a particular MOS and Zone are eligible for reenlistment due to MOS inventory limitations and other constraints. SRB lump sum payments are calculated by multiplying the Marine's monthly basic pay at the time of reenlistment by the number of years (partial years included) and the SRB multiple. (Marine Corps Order 7220.24M, 1990) The following figures show the multiples offered for FY2005 reenlistment.



| | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| COUNT | 11048 | 5250 | 780 | 4286 | 400 | 109 |
| PERCENT | 50.5 | 24.0 | 3.6 | 19.6 | 1.8 | 0.5 |

Figure 17. FY2005 first-term SRB Multiples.



|  | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| COUNT | 13589 | 1442 | 284 | 514 | 43 | 8 |
| PERCENT | 85.6 | 9.1 | 1.8 | 3.2 | 0.3 | 0.1 |

Figure 18. FY2005 subsequent-term SRB Multiples.

#### *c. PMOS*

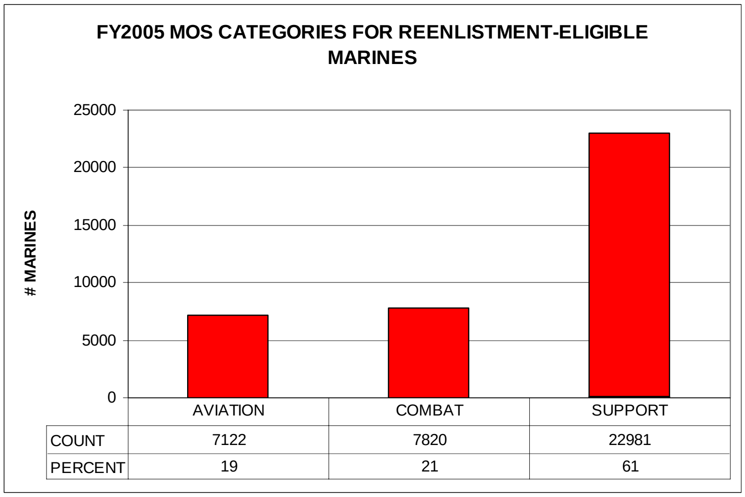
PMOS represents the Marines Primary Military Occupational Specialty. There are over 200 of these in the Marine Corps, each represented by a four-digit code. Examples used in the next chapter include Infantry Rifleman (0311), Warehouse Clerk (3051), and CH-53E Helicopter Mechanic (6113).

#### *d. OCCFIELD*

OCCFIELDs consist of all PMOSs that have the same first two digits. For instance, the CH-53E Helicopter Mechanic (PMOS 6113) and CH-53E Crew Chief (6173) both fall within the 61 OCCFIELD. There are over 30 OCCFIELDs in the Marine Corps.

#### *e. MOSCAT*

Due to computational constraints, the "tree" method in S-Plus limits categorical variables to a maximum of 24 levels; therefore, neither PMOS nor OCCFIELD variables could be used without collapsing them somehow. The MOSCAT variable simply groups all OCCFIELDs into the categories of "Combat," "Aviation," and "Support." This is, admittedly, a crude way to group OCCFIELDs, since an Administrative Clerk and Diesel Mechanic both fall within the "Support" category, and they are very different occupations. Future work might entail finding a better method for grouping these occupations that will provide more detail in modeling. Figure 19 shows the MOSCAT categories for combined first- and subsequent-term populations.

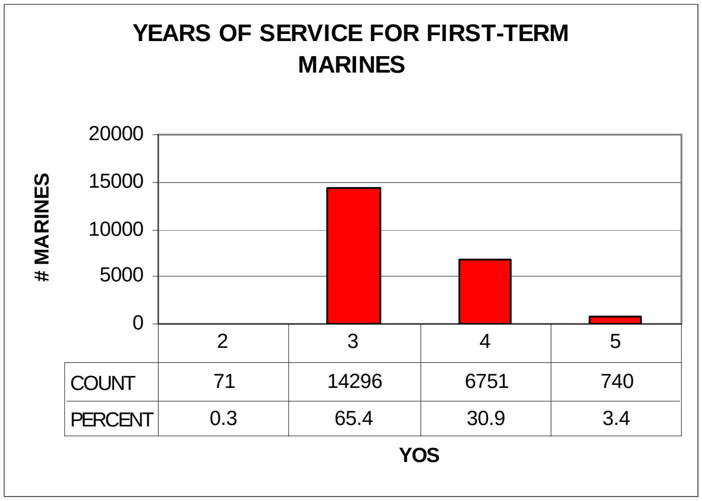


|  | AVIATION | COMBAT | SUPPORT |
|---|---|---|---|
| COUNT | 7122 | 7820 | 22981 |
| PERCENT | 19 | 21 | 61 |

Figure 19. MOS Categories for reenlistment eligible population, end FY2004.

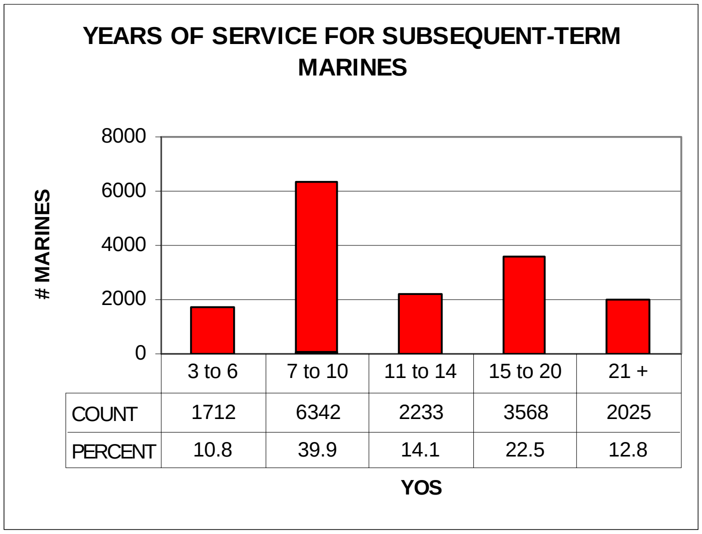
### *f. YOS (Years of Service)*

The YOS variable indicates the of years of service a Marine has on the last day of FY2004, which is virtually the first day of FY2005, the year in which he or she was eligible to reenlist.



|  | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| COUNT | 71 | 14296 | 6751 | 740 |
| PERCENT | 0.3 | 65.4 | 30.9 | 3.4 |

Figure 20. YOS for first-term reenlistment population, end FY2004.



| | 3 to 6 | 7 to 10 | 11 to 14 | 15 to 20 | 21 + |
|---|---|---|---|---|---|
| COUNT | 1712 | 6342 | 2233 | 3568 | 2025 |
| PERCENT | 10.8 | 39.9 | 14.1 | 22.5 | 12.8 |

Figure 21. YOS for subsequent-term reenlistment population, end FY2004.

THIS PAGE INTENTIONALLY LEFT BLANK

# V. ANALYSIS AND RESULTS

To review, the intent of this research was to construct two logistic regression models to predict reenlistment of Marines in FY 2005, one model for the first-term Marine population and one for the subsequent-term population. In order to do this, the models were fit using data from the 2004 reenlistment population. The reason for doing this is to use the most recent information to characterize current trends in the prediction of the next year's reenlistments. Essentially such an approach assumes that the current year requiring prediction is much like the previous year. This is in contrast to work in the literature reviewed, which either focuses on only one year's worth of data, or groups many years together, to explore the significant factors in reenlistment.

After first-term and subsequent-term Marines were separated into two data sets, the available variables were explored to determine which ones might contribute the most to predicting reenlistment. Classification trees were grown, cross-validated, and pruned in order to determine the optimal tree size with the best predictive power. This was done separately for the first- and subsequent-term Marine cohorts with enlistment contracts ending in FY2004. In looking at trees from FY2001 to FY2004, first-term reenlistment predictions (classified by ELIGREEN1 = 0 for "was not retained" and ELIGREEN1 = 1 for "retained") did not achieve better than roughly 70 percent correct classifications. Consistently, subsequent-term predictions were slightly better, at around 75 percent correct classification.

It is interesting to note that the classification trees were not able to reach a high level of predictive power. This may indicate that there are important factors that affect reenlistment that are not being captured by the current set of predictors. In any case, CART was not employed to do predictions but rather to provide insight into which variables might be useful in logistic regression. Further, the trees were also helpful for suggesting possible modifications to the variables, such as collapsing many levels into few, or making numerical variables into categorical variables.

## A. FIRST-TERM MODELS

Figure 22 is the cross-validation plot used in creating the tree for first-term reenlistment predictions. The horizontal axis represents the size of the tree and the vertical axis the number of misclassified instances. Notice that a large tree of, say, 100 nodes (depicted on the *x*-axis), does not predict much better than a tree of smaller size – especially when considering the data set has 22,655 observations. That is, note that the *y*-axis, which is the number of misclassifications, has a range from just under 6,780 to just under 6,900, for a range of about 220 misclassifications, or roughly one percent of the total number of observations. So, the largest tree with over 100 nodes does only very slightly better than the smallest trees of just one or two nodes, and the best tree in terms of misclassification rate has about 30 nodes.

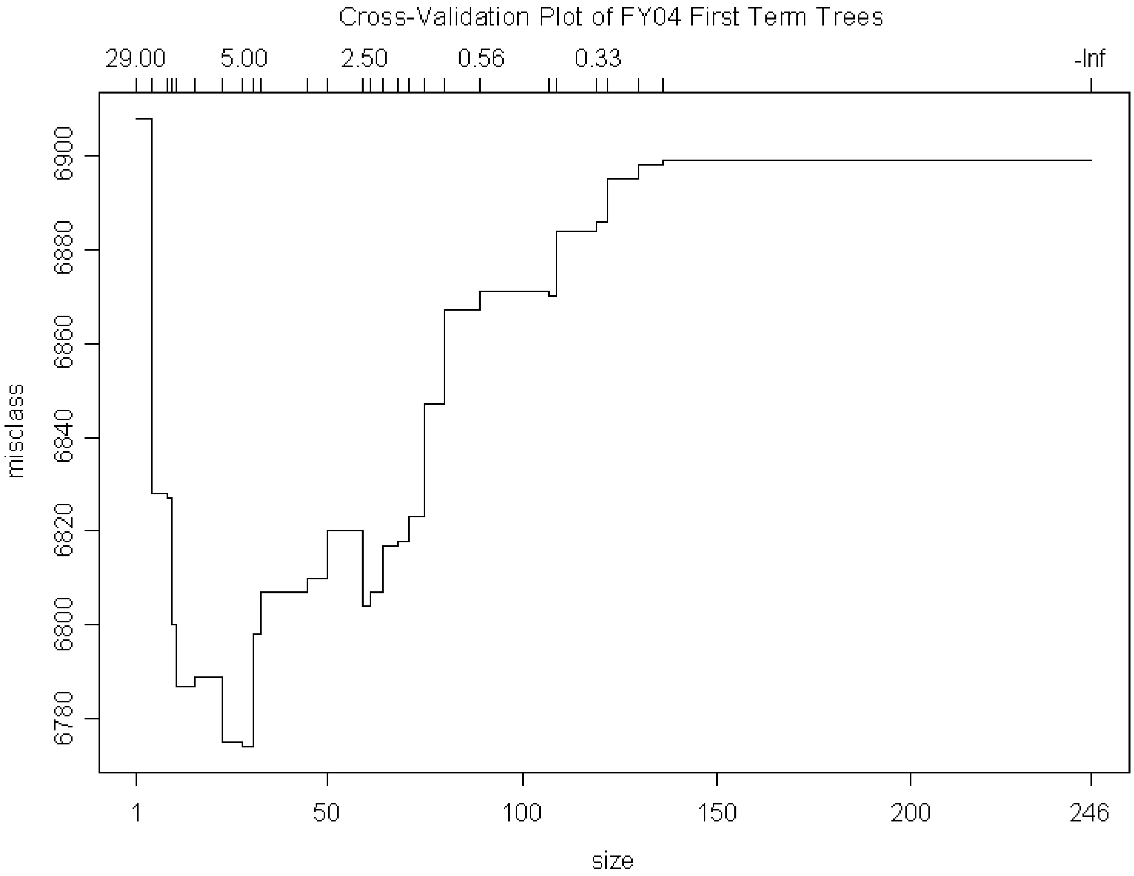


Figure 22. Cross-validation plot of first-term classification tree.

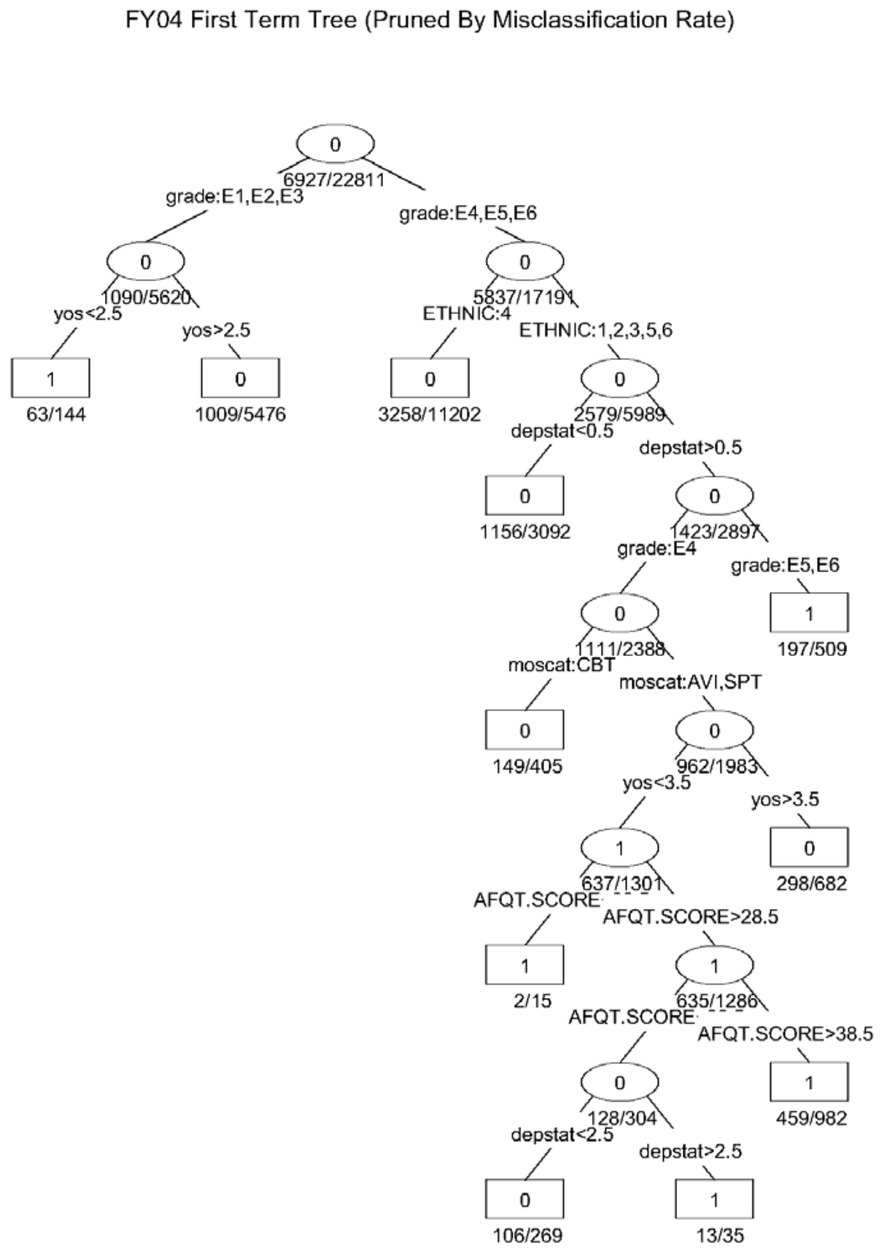


Figure 23. Classification tree for FY2004 first-term reenlistments.

Figure 23 shows the tree used for developing the first-term logistic regression model. Recall that FY2004 data was used in developing both the tree and model for predicting FY2005 reenlistment. The tree shown above predicts reenlistment based on the same data from which it was created. The first attempted logistic regression model used only the first three variables chosen by the classification tree: GRADE, YOS, and ETHNIC. This proved to be one of the better prediction models for the first-term population.

It was not surprising that, prior to using the classification tree as an aid in variable selection, the fitting of early logistic regression models resulted in the inclusion of many of the best predictor variables identified with the tree. However, we then used the information from the tree to better define the logistic regression predictors. For example, since the tree had the variable YOS split at YOS > 2.5, a categorical variable having two levels of YOS was formed (less than three YOS and three or more YOS). Other categorical variables, based on the tree splits, were also used in seeking a better model than the original logistic regression “winner.” ETHNIC was collapsed into two levels instead of six. Using these two, new categorical variables improved the prediction error, $\text{AVGDIFF}_{\text{MODEL}}$, from 585 to 555.4. Other categorical variables, listed in Table 4, were created and placed into the logistic regression, but they did not prove useful in lowering $\text{AVGDIFF}_{\text{MODEL}}$.

Table 4. Categorical variables created based on classification tree.

| Variable | Definition |
|---|---|
| HIGRADE | {E1,E2,E3} OR {E4,E5,E6} |
| ETHI | {WHITE} OR {ALL OTHER} |
| DEPI | {NO DEPENDENTS} OR {HAS DEPENDENTS} |
| CBTMOS | {COMBAT MOS} OR {AVIATION OR SUPPORT} |
| AFQTI | {AFQT >= 29} OR {AFQT < 29} |
| YOSI | {YOS >= 3} OR {YOS < 3} END OF FY PRIOR TO ECC |

*Note: all new categorical variables have two levels, based on binary splits in tree. Such modifications proved useful in this case, but limiting to two levels is not required. YOSI and ETHI were the two useful variables formed based on classification tree results.*

Table 5 summarizes several of the variable sets used and provides a comparison of performance based on $\text{AVGDIFF}_{\text{MODEL}}$. Model ‘M’ was the best model for first-term prediction. Surprisingly, including SRBELIG in the models did not decrease the prediction error. Given the CNA results discussed in Chapter II, we also tried a similar interaction of SRBELIG and AFQTI, but it too did not prove helpful in decreasing the prediction error. While this result is consistent with the results from the pruned classification tree (SRBELIG was not in the variables included by the `tree` method after cross-validation and pruning) it is nonetheless surprising and warrants further study.

Since the classification tree had splits for GRADE in more than one instance, a categorical variable was made with four levels to match these partitions: GRADE = {E1, E2, E3}, GRADE = {E4}, and GRADE = {E5, E6}. This did not improve the prediction error achieved in logistic regression, as measured by AVGDIFF$_{\text{MODEL}}$. Therefore, the original GRADE variable with six levels (when considering the first-term population) was used.

Table 5. Summary of variable selection and AVGDIFF$_{\text{MODEL.}}$

| MODEL | AVDIFF | VARIABLES USED |
|---|---|---|
| A | 599.5 | ETHNIC DEPSTAT GRADE SRBELIG OCCFIELD SEX MARSTAT AFQT_SCORE YOS |
| B | 585.0 | GRADE YOS ETHNIC |
| C | 590.4 | GRADE YOS ETHNIC DEPSTAT |
| D | 711.1 | HIGRADE YOS ETHNIC |
| E | 587.3 | GRADE YOS ETHI |
| F | 584.5 | GRADE YOS ETHNIC DEPI |
| G | 634.8 | GRADE YOS ETHNIC CBTMOS |
| H | 645.9 | GRADE YOS ETHNIC SRBELIG |
| I | 650.6 | GRADE YOS ETHNIC SRBELIG AFQTI SRBELIG\|AFQTI |
| J | 2418.9 | GRADE YOS ETHNIC SRBELIG MOSCAT SRBELIG\|MOSCAT |
| K | 557.3 | GRADE YOSI ETHNIC |
| L | 676.6 | HIGRADE YOSI ETHNIC |
| M | 555.4 | GRADE YOSI ETHI |

*Not all variable selection sets used in finding the best model are shown here.*

A description of the model fit at some level below the overall fit is warranted. To accomplish this, three MOSs were selected for comparison at the GRADE = E4 level. These MOSs were selected due to the author's familiarity and because each one fits into a different category as described by the variable MOSCAT – a convenient classification. MOS 0311 (Rifleman), MOS 3051 (Warehouse Clerk), and MOS 6113 (CH-53 Helicopter Mechanic), were selected, and fit the Combat, Aviation, and Support categories, respectively. (Marine Corps Order P1200.16, 2005) The following tables depict model performance as it relates to predicted numbers of reenlistments compared to actual numbers of reenlistments.

Table 6. MOS 0311 model performance comparison.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 1590 | 556.97 | 466.00 | 90.97 | 8275.48 | 14.86 |
| B | 1590 | 505.90 | 466.00 | 39.90 | 1591.84 | 3.15 |
| M | 1590 | 489.19 | 466.00 | 23.19 | 537.91 | 1.10 |

Table 7. MOS 3051 model performance comparison.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 296 | 136.59 | 118.00 | 18.59 | 345.70 | 2.53 |
| B | 296 | 114.51 | 118.00 | -3.49 | 12.17 | 0.11 |
| M | 296 | 113.31 | 118.00 | -4.69 | 22.04 | 0.19 |

Table 8. MOS 6113 model performance comparison.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 26 | 7.92 | 5.00 | 2.92 | 8.50 | 1.07 |
| B | 26 | 6.24 | 5.00 | 1.24 | 1.55 | 0.25 |
| M | 26 | 7.60 | 5.00 | 2.60 | 6.75 | 0.89 |

In Tables 6–8, the “ELIGIBLE” column shows the number Marines eligible for reenlistment during FY2005 in each MOS, who were of GRADE E4. The right-most column, AVGDIFF, shows that prediction cell’s contribution to the overall lack of fit of the model, $\text{AVGDIFF}_{\text{MODEL}}$. It is evident in the tables that no single model dominates the others when comparing within the selected groups of Marines. However, the original measure of performance, $\text{AVGDIFF}_{\text{MODEL}}$, is very useful in determining a “winner.” In Figure 24, shown on the next page, the best is Model M.

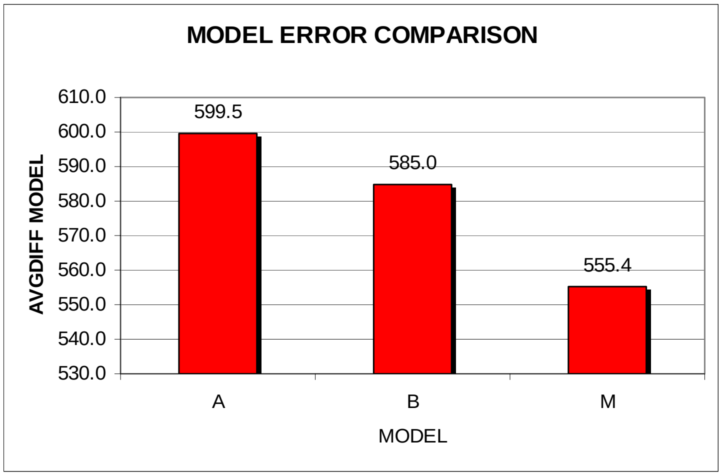


Figure 24. Comparison of model error (measured by $\text{AVGDIFF}_{\text{MODEL}}$).

Model M has the lowest prediction error when taking into account the predictions over all MOS and GRADE cells. What is most interesting is that adding more terms in the model, contrary to what one might intuitively think, does not improve predictive power. Rather, it seems to introduce additional "noise" into the predictions, actually degrading model performance.

## B. SUBSEQUENT-TERM MODELS

To find a reasonable model to predict subsequent-term reenlistments, the same steps were followed as used in modeling first-term population reenlistments. The classification tree is shown on the next page, in Figure 25.

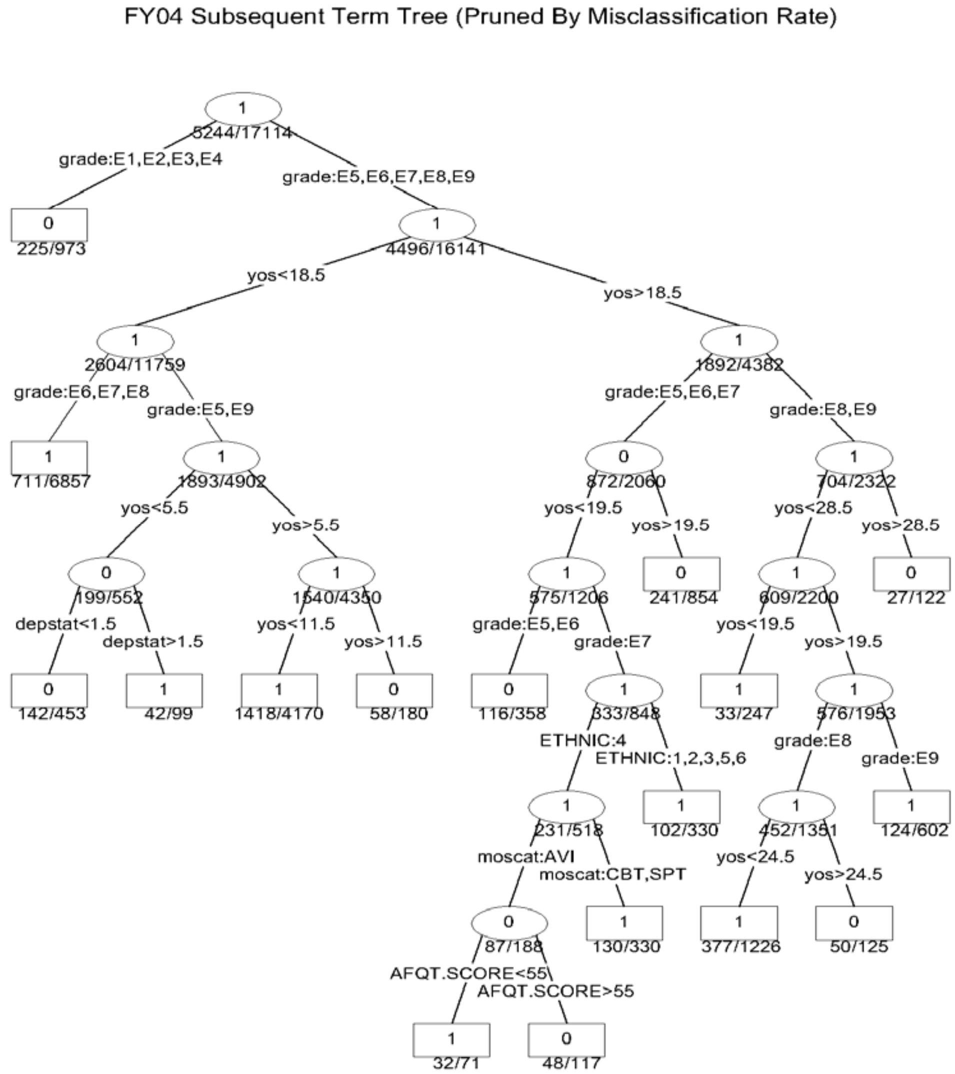


Figure 25. Classification tree for FY2004 subsequent-term reenlistments.

In logistic regression for the subsequent-term population, using the first three variables selected for partitioning from the classification tree did not prove useful as it did in model-building for the first-term population, based on the AVGDIFF$_{\text{MODEL}}$ score. However, using information from the tree’s variable splits did improve the score. Useful categorical variables were formed based on the classification tree partitions of the variables GRADE and ETHNIC.

GRADE was collapsed from nine levels to four, based on the left side of the tree. The first split on the variable GRADE separated the lowest four levels from the remaining five levels. After this, the groups of GRADE = {E6, E7, E8} and GRADE = {E5, E9} were partitioned. The grouping of E5 and E9 together in a node provided no useful, intuitive value. Therefore, E5 and E9 were given separate levels in the newly formed categorical value based on GRADE (called GLEVEL). See Table 9 for a summary of variables formed based on the classification tree.

The tree's second split used the variable YOS. Categorical variables with three and five levels were utilized and compared. The version with three levels proved more effective, but neither was better than using YOS in its original form. DEPSTAT was grouped into two levels, with comparison of the thresholds at DEPSTAT ≥ 1 and DEPSTAT ≥ 2, with the neither version improving predictions. Utilizing the splits on the variables AFQT and MOSCAT did not improve results. This is not surprising because they are split relatively late (or low) in the structure of the tree.

Table 9. Categorical variables created based on classification tree.

| Variable | Definition |
|---|---|
| GLEVEL | {E1,E2,E3,E4} OR {E5} OR {E6,E7,E8} OR {E9} |
| ETHI | {WHITE} OR {ALL OTHER} |
| DEPI | {>= 1 DEPENDENT} OR {0 DEPENDENTS} |
| AVIMOS | {AVIATION MOS} OR {COMBAT OR SUPPORT} |
| AFQTI | {AFQT >= 55} OR {AFQT < 55} |
| YOSI | {YOS > 18} OR { 6 <= YOS =< 18} OR {YOS < 6} |

Remarkably, the same partition on ETHNIC occurred in the subsequent-term tree for as for the first-term tree. This split was utilized and improved the model's score once again. Below, Table 10 summarizes the results from several of the attempted models for comparison. Models C and G are highlighted to show that they are two of the better results, with very similar overall error measurements. These models are the lowest in the AVDIFF column. There are two differences between Model C here and Model M from the first-term set. First, in subsequent-term Model C, GLEVEL replaced GRADE, which was kept in its original format for the first term. Second, YOS was left in its original form because it provided better results than the categorical version, YOSI. Model G, which is the same as Model C with SRBELIG added, improved the overall fit only slightly. Reaching a limited improvement was not surprising since SRBELIG did not appear once in the pruned classification tree. Using an interaction between SRBELIG and AFQT_SCORE as suggested in the reviewed literature did not improve model performance.

Table 10. Summary of variable selection and $AVGDIFF_{MODEL.}$

| MODEL | AVGDIFF | VARIABLES USED |
|---|---|---|
| A | 733.2 | ETHNIC DEPSTAT GRADE SRBELIG OCCFIELD SEX MARSTAT AFQT_SCORE YOS |
| B | 720.4 | OCCFIELD GRADE ETHNIC |
| C | 693.7 | YOS GLEVEL ETHI |
| D | 703.2 | YOSI GLEVEL ETHI |
| E | 732.3 | OCCFIELD GLEVEL ETHI |
| F | 694.2 | YOS GLEVEL ETHI DEPI |
| G | 693.3 | YOS GLEVEL ETHI SRBELIG |
| H | 695.3 | YOS GLEVEL ETHI SRBELIG AFQT_SCORE SRBELIG\|AFQT_SCORE |
| I | 696.3 | YOS GLEVEL ETHI SRBELIG AFQTI SRBELIG\|AFQTI |
| J | 698.5 | YOS GLEVEL ETHI MOSCAT |
| K | 700.4 | YOS GLEVEL ETHI AVIMOS |

To show results based on specific MOSs and GRADE = E6 for the subsequent-term population, MOS 0369 (Infantry Unit Leader) replaces MOS 0311 from the first-term example. The remaining two MOS designations, 3051 and 6113, are used again here, except at the E6 GRADE level. MOS 0369 replaces MOS 0311 because MOS 0311 is primarily a first-term Marine MOS and 0369 is only filled by subsequent-term Marines.

Table 11. MOS 0369 model performance comparison.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 405.00 | 315.09 | 314.00 | 1.09 | 1.19 | 0.00 |
| C | 405.00 | 337.23 | 314.00 | 23.23 | 539.67 | 1.60 |
| G | 405.00 | 338.37 | 314.00 | 24.37 | 594.09 | 1.76 |

Table 12. MOS 3051 model performance comparison.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 72.00 | 57.83 | 66.00 | -8.17 | 66.72 | 1.15 |
| C | 72.00 | 60.26 | 66.00 | -5.74 | 32.98 | 0.55 |
| G | 72.00 | 60.42 | 66.00 | -5.58 | 31.14 | 0.52 |

Table 13. MOS 6113 model performance comparison.

| MODEL | ELIGIBLE | PREDICTED | ACTUAL | DIFFERENCE | SQ DIFF | AVG DIFF |
|---|---|---|---|---|---|---|
| A | 21.00 | 18.51 | 17.00 | 1.51 | 2.29 | 0.12 |
| C | 21.00 | 18.74 | 17.00 | 1.74 | 3.04 | 0.16 |
| G | 21.00 | 18.74 | 17.00 | 1.74 | 3.03 | 0.16 |

The overall model comparisons are shown below, using $\text{AVGDIFF}_{\text{MODEL}}$.

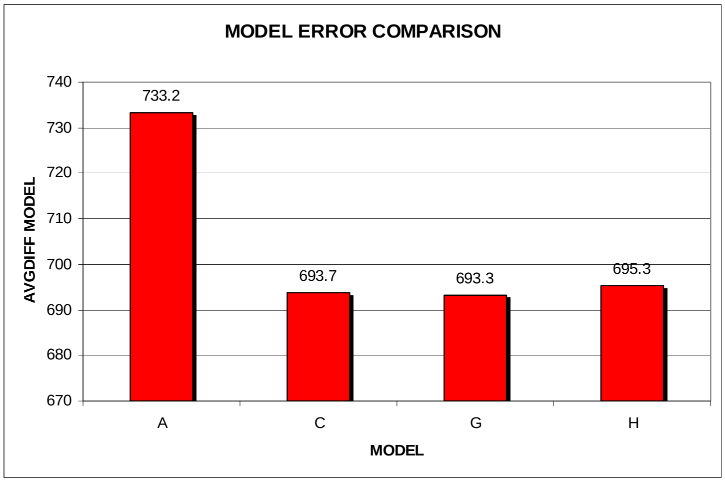


Figure 26. Comparison of model error (measured by $\text{AVGDIFF}_{\text{MODEL}}$).

In relation to the first-term population results, a similar observation can be made here about the error scores of the models. Neither of the overall winners, Models C and G (which virtually tied), dominated when compared within the three selected MOSs at GRADE = E6. Once again, the priority here was on overall model fit, as measured by the $\text{AVGDIFF}_{\text{MODEL}}$ statistic.

THIS PAGE INTENTIONALLY LEFT BLANK

# VI. CONCLUSIONS AND RECOMMENDATIONS

## A. CONCLUSIONS

In creating logistic regression models to forecast reenlistments of first- and subsequent-term Marines, classification trees were used to determine what variables may be important predictors. Generally, the first few partitions defined in the trees proved useful in taking past data, and applying insight gained in the prediction of the next year's reenlistments. Not all attempts to create new variables based on the trees' binary splits improved the models' error score, but evidence of improvement was shown by re-coding variables such as years of service, grade, and the race and ethnic codes.

Logistic regression provided varying results in predicting reenlistments across the many cells of Marines, indexed by MOS and GRADE. No one model dominated in the reenlistment predictions for each selected MOS and GRADE. Therefore, the "goodness of fit" measurement provided a useful means to compare the overall performance of the models. In seeking the best goodness of fit, the variables were remarkably similar between the first-term and subsequent-term models. Future work will include finding ways to reduce prediction errors to meet an acceptable standard, dictated by the Marine Corps.

The results in Chapter V are a beginning to the efforts to develop the Career Force Retention Model desired by M&RA. Combined with other continued efforts, the ability to predict force inventory and structure will continue to develop. A thesis entitled *Determining the Number of Reenlistments Necessary to Satisfy Future Force Requirements*, by Captain J. David Raymond (2006), forecasts changes in the population of Marines not eligible for promotion in a fiscal year. These forecasts are based on promotion and attrition rates and MOS changes. Combining the prediction output of this thesis and Raymond's work, results in a prediction of the next year's enlisted force.

## B. RECOMMENDATIONS FOR FUTURE WORK

The inclusion of deployment data from TFDW should be one of the first steps in continuing work to improve reenlistment predictions. This data is available, and can be readily merged into the existing longitudinal data set. SAS code has been written for transforming the transactional format of this data for use with the previously developed longitudinal data set. Factors such as the frequency of deployments and number of deployed days for each year could prove to be useful in predicting reenlistment. In addition to deployment data, exit survey data might be used to determine reasons for attrition from the Marine Corps. Exit surveys can uncover key factors involved in the reenlistment decision, and may be useful in determining variables needed in forecasting reenlistment.

As described in Chapter I, first-term Marines' reenlistments may be limited in certain MOSs due to force structure constraints. The amount of reenlistments can be limited by the amount of boat spaces available. Future work should explore a method that accounts for these constraints.

# LIST OF REFERENCES

Delany, K.S.. An *Analysis of Factors that Influence Reenlistment Decisions in the U.S. Army*, M.S. Thesis, Naval Postgraduate School, Monterey, California, 1997.

Devore, J.L. (2004) *Probability and Statistics for Engineering and the Sciences*. Toronto, Ontario, Canada: Brooks/Cole – Thomson.

Fleiss, J.L., Levin, B. & Paik, M.C. (2003) *Statistical Methods for Rates and Proportions*. Hoboken, NJ: John Wiley & Sons.

Hand, D., Mannila, H. & Smyth, P. (2001) *Principles of Data Mining*. Cambridge, MA: The MIT Press.

Hattiangadi, A.U., Ackerman, D., Kimble, T. & Quester, A.O. (2004) *Cost-Benefit Analysis of Lump Sum Zone A, Zone B, and Zone C Reenlistment Study: Final Report*. (CRM D0009652A2). Alexandria, VA: The Center for Naval Analyses.

Hinson, William B., *A Statistical Analysis of Individual Success after Successful Completion of Defense Language Institute Foreign Language Center Training,* M.S. Thesis, Naval Postgraduate School, Monterey, California, 2005.

North, J. and Quester, A. (1992). *Determining the Number and Composition of First-Term Reenlistments: The first term alignment plan (FTAP).* (CRM 92-85). Alexandria, VA: The Center for Naval Analyses.

Purcell, T.S. *The Use of Classification Trees to Characterize the Attrition Process for Army Manpower Models*, M.S. Thesis, Naval Postgraduate School, Monterey, California, 1999.

Quester, A. and North, J. (1993). *Managing the Enlisted Marine Corps in the 1990s Study: Final Report*. (CRM 92-202). Alexandria, VA: The Center for Naval Analyses.

Raymond, J.D. *Determining the Number of Reenlistments Necessary to Satisfy Future Force Requirements*, M.S. Thesis, Naval Postgraduate School, Monterey, California, 2006.

United States Government Accountability Office. (2005) *DoD Needs Action Plan to Address Enlisted Personnel Recruitment and Retention Challenges*. Washington, D.C.

United States Marine Corps. (2004) *Enlisted Force Career Planning and Retention Manual* (*Marine Corps Order P1040.31J).*

United States Marine Corps. (2005) *Marine Occupational Specialties (MOS) Manual (Marine Corps Order P1200.16).*

United States Marine Corps. (1990) *Selective Reenlistment Bonus (SRB) Program (Marine Corps Order 7220.24M).*

Zamarippa, L. R. (2005). *Manpower 101*. Microsoft PowerPoint Presentation. Headquarters, Marine Corps, Manpower and Reserve Affairs.

# APPENDIX A. SAS CODE

This appendix contains the SAS code used in assembling the longitudinal data set, and for extracting subsets of it for use in making classification trees. The code used for logistic regression in SAS is shown last.

```
LIBNAME Demo 'Z:\Demogr';
LIBNAME Long 'Z:\Temp';

option YEARCUTOFF = 1950;

*IMPORT DBF FILES TO SAS DATA SETS.  UPDATE THE LIST BELOW OF YEARS TO
RUN 'GATHER' MACRO;

%let LIST = 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
2000 2001 2002 2003 2004 2005;
%MACRO GATHER;
%DO I= 1 %TO 18;
%LET YR = %SCAN(&LIST, &I);
PROC IMPORT OUT = Long.data&YR
                  DATAFILE = "Z:\Demogr\FY&YR..dbf"
                  DBMS=DBF REPLACE;
                  GETDELETED = NO;
RUN;
PROC SORT DATA = Long.data&YR OUT = Long.sort&YR NODUPKEY;
      BY SSN;
RUN;
%END;
;
%MEND;
%GATHER;
```

```
* USE MACRO 'NAMER' (WITH UPDATED LIST) TO RENAME THE TFDW FIELDS BY
YEAR AND DROP UNWANTED FIELDS;
%let LIST = 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
2000 2001 2002 2003 2004 2005;
%MACRO NAMER;
%DO I = 1 %TO 18;
%LET YR = %SCAN(&LIST, &I);
DATA Long.refine&YR (rename=(PRESENT_GR=GRADE&YR OCCFIELD=OCC&YR
      PRIMARY_MO=PMOS&YR ECC_EAS_FL=ECCFL&YR EXPIRATIO2=ECC&YR
      DUTY_STATU=DUST&YR RECORD_STA=REC&YR MARITAL_ST=MAR&YR
      NUM_DEPEND=DEP&YR YOS=YOS&YR CURRENT_SO=SOURCE&YR
      CURRENT_EN=ENLN&YR CRISIS_COD=CCODE&YR CRISIS_PAR=CDATE&YR
      EXPIRATION=EAS&YR PLANNED_RE=PLAN&YR PLANNED_R2=PLANFL&YR
      SELECTIVE_=ZONE&YR INITIAL_AC=IADD&YR PAY_ENTRY_=PEBD&YR));

     SET Long.sort&YR (DROP = PRESENT_RE PRIOR_CONT PROFICIENC
      PROFICIEN2 PROFICIEN3 REENLISTME PHYSICAL_F PHYSICAL_2 PRIOR_PHYS
      PRIOR_PHY2 WEIGHT_CON ADDL_FIRST ADDL_SECON COMPONENT_ STRENGTH_C
      PLANNED_R3 PLANNED_R4 GRADE_SELE LAST_NAME FIRST_NAME BILLET_MOS
      CURRENT_AC GEOGRAPHIC GEOGRAPHI2 PRESENT_MO);

RUN;
      %END;

%MEND NAMER;
%NAMER;

* USE MACRO 'NAMER' (WITH UPDATED LIST) TO RENAME THE TFDW FIELDS BY
YEAR AND DROP UNWANTED FIELDS;
%let LIST = 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999
2000 2001 2002 2003 2004 2005;
%MACRO JOINEMUP;
DATA Demo.joinem;

      MERGE %DO J=1 %TO 18;
                  %LET YR = %SCAN(&LIST, &J);
                  Long.refine&YR (in=Indata&YR)
                  %END;
      ;
      BY SSN;
RUN;
%MEND JOINEMUP;
%JOINEMUP;


DATA Demo.joinem3;
      SET Demo.joinem;

lastday1988 = '30sep1988'D;
lastday1989 = '30sep1989'D;
lastday1990 = '30sep1990'D;
lastday1991 = '30sep1991'D;
lastday1992 = '30sep1992'D;
lastday1993 = '30sep1993'D;
lastday1994 = '30sep1994'D;
lastday1995 = '30sep1995'D;
```

```
lastday1996 = '30sep1996'D;
lastday1997 = '30sep1997'D;
lastday1998 = '30sep1998'D;
lastday1999 = '30sep1999'D;
lastday2000 = '30sep2000'D;
lastday2001 = '30sep2001'D;
lastday2002 = '30sep2002'D;
lastday2003 = '30sep2003'D;
lastday2004 = '30sep2004'D;
lastday2005 = '30sep2005'D;

ARRAY ELIGREEN1[*]  ELIGREEN11989 ELIGREEN11990 ELIGREEN11991
      ELIGREEN11992 ELIGREEN11993 ELIGREEN11994 ELIGREEN11995
      ELIGREEN11996 ELIGREEN11997 ELIGREEN11998 ELIGREEN11999
      ELIGREEN12000 ELIGREEN12001 ELIGREEN12002
      ELIGREEN12003 ELIGREEN12004 ELIGREEN12005;

ARRAY MONTHSVC[*]  MSVC2000 MSVC2001 MSVC2002 MSVC2003 MSVC2004;

ARRAY LASTDAY[*]   lastday2000 lastday2001 lastday2002 lastday2003
                   lastday2004;

ARRAY ECC[*]      ECC1988-ECC2005;

ARRAY ECCFY[*]    ECCFY1989 ECCFY1990 ECCFY1991 ECCFY1992 ECCFY1993
                  ECCFY1994 ECCFY1995 ECCFY1996 ECCFY1997 ECCFY1998
                  ECCFY1999 ECCFY2000 ECCFY2001 ECCFY2002 ECCFY2003
                  ECCFY2004 ECCFY2005;

ARRAY PMOSX[*]    PMOS1988-PMOS2005;

ARRAY GRADE[*]    GRADE1988-GRADE2005;

ARRAY NEWBIE[*]   NEWBIE1989 NEWBIE1990 NEWBIE1991 NEWBIE1992
                  NEWBIE1993 NEWBIE1994 NEWBIE1995 NEWBIE1996
                  NEWBIE1997 NEWBIE1998 NEWBIE1999 NEWBIE2000
                  NEWBIE2001 NEWBIE2002 NEWBIE2003 NEWBIE2004
                  NEWBIE2005;

ARRAY LOSS[*]     LOSS1989 LOSS1990 LOSS1991 LOSS1992 LOSS1993 LOSS1994
                  LOSS1995 LOSS1996 LOSS1997 LOSS1998 LOSS1999 LOSS2000
                  LOSS2001 LOSS2002 LOSS2003 LOSS2004 LOSS2005;

ARRAY TRANSITION[*] $15. TRANSITION1989 TRANSITION1990 TRANSITION1991
                         TRANSITION1992 TRANSITION1993 TRANSITION1994
                         TRANSITION1995 TRANSITION1996 TRANSITION1997
                         TRANSITION1998 TRANSITION1999 TRANSITION2000
                         TRANSITION2001 TRANSITION2002 TRANSITION2003
                         TRANSITION2004 TRANSITION2005;

ARRAY MOSGRADE[*] $6. MOSGRADE1989 MOSGRADE1990 MOSGRADE1991
                      MOSGRADE1992 MOSGRADE1993 MOSGRADE1994
                      MOSGRADE1995 MOSGRADE1996 MOSGRADE1997
                      MOSGRADE1998 MOSGRADE1999 MOSGRADE2000
                      MOSGRADE2001 MOSGRADE2002 MOSGRADE2003
                      MOSGRADE2004 MOSGRADE2005;
```

```
ARRAY ECCFYTEST[*] ECCFYTEST1989 ECCFYTEST1990 ECCFYTEST1991
                   ECCFYTEST1992 ECCFYTEST1993 ECCFYTEST1994
                   ECCFYTEST1995 ECCFYTEST1996 ECCFYTEST1997
                   ECCFYTEST1998 ECCFYTEST1999 ECCFYTEST2000
                   ECCFYTEST2001 ECCFYTEST2002 ECCFYTEST2003
                   ECCFYTEST2004 ECCFYTEST2005;

ARRAY OCC[*] $2.   OCC1997 OCC1998 OCC1999 OCC2000 OCC2001 OCC2002
                   OCC2003 OCC2004 OCC2005;

ARRAY MOSCAT[*]   $3.   MOSCAT1997 MOSCAT1998 MOSCAT1999 MOSCAT2000
                        MOSCAT2001 MOSCAT2002 MOSCAT2003 MOSCAT2004
                        MOSCAT2005;

ARRAY SRBZ[*] $1. SRBZONE2001 SRBZONE2002 SRBZONE2003 SRBZONE2004
SRBZONE2005;

DO K = 1 TO 17;
      IF ECC[K] > lastday1988 AND ECC[K]<= lastday1989 THEN DO;
            ECCFY1989 = 1; ECCFYTEST1989 = 1; END;
      IF ECC[K] > lastday1989 AND ECC[K]<= lastday1990 THEN DO;
            ECCFY1990 = 1; ECCFYTEST1990 = 1; END;
      IF ECC[K] > lastday1990 AND ECC[K]<= lastday1991 THEN DO;
            ECCFY1991 = 1; ECCFYTEST1991 = 1; END;
      IF ECC[K] > lastday1991 AND ECC[K]<= lastday1992 THEN DO;
            ECCFY1992 = 1; ECCFYTEST1992 = 1; END;
      IF ECC[K] > lastday1992 AND ECC[K]<= lastday1993 THEN DO;
            ECCFY1993 = 1; ECCFYTEST1993 = 1; END;
      IF ECC[K] > lastday1993 AND ECC[K]<= lastday1994 THEN DO;
            ECCFY1994 = 1; ECCFYTEST1994 = 1; END;
      IF ECC[K] > lastday1994 AND ECC[K]<= lastday1995 THEN DO;
            ECCFY1995 = 1; ECCFYTEST1995 = 1; END;
      IF ECC[K] > lastday1995 AND ECC[K]<= lastday1996 THEN DO;
            ECCFY1996 = 1; ECCFYTEST1996 = 1; END;
      IF ECC[K] > lastday1996 AND ECC[K]<= lastday1997 THEN DO;
            ECCFY1997 = 1; ECCFYTEST1997 = 1; END;
      IF ECC[K] > lastday1997 AND ECC[K]<= lastday1998 THEN DO;
            ECCFY1998 = 1; ECCFYTEST1998 = 1; END;
      IF ECC[K] > lastday1998 AND ECC[K]<= lastday1999 THEN DO;
            ECCFY1999 = 1; ECCFYTEST1999 = 1; END;
      IF ECC[K] > lastday1999 AND ECC[K]<= lastday2000 THEN DO;
            ECCFY2000 = 1; ECCFYTEST2000 = 1; END;
      IF ECC[K] > lastday2000 AND ECC[K]<= lastday2001 THEN DO;
            ECCFY2001 = 1; ECCFYTEST2001 = 1; END;
      IF ECC[K] > lastday2001 AND ECC[K]<= lastday2002 THEN DO;
            ECCFY2002 = 1; ECCFYTEST2002 = 1; END;
      IF ECC[K] > lastday2002 AND ECC[K]<= lastday2003 THEN DO;
            ECCFY2003 = 1; ECCFYTEST2003 = 1; END;
      IF ECC[K] > lastday2003 AND ECC[K]<= lastday2004 THEN DO;
            ECCFY2004 = 1; ECCFYTEST2004 = 1; END;
      IF ECC[K] > lastday2004 AND ECC[K]<= lastday2005 THEN DO;
            ECCFY2005 = 1; ECCFYTEST2005 = 1; END;
END;

DO L = 1 TO 17;
      IF ECCFY[L]=. AND PMOSX[L+1]~='' THEN ECCFY[L]=0;
      IF ECCFYTEST[L]=. THEN ECCFYTEST[L]=0;
```

```
END;

DO M = 1 TO 17;
********CAPTURES ELIGIBILITY AND EENLISTMENT (LOOKING AT END OF FY THAT
ECC IS IN);
      IF ECCFY[M] = 1 AND PMOSX[M+1]~='' THEN ELIGREEN1[M] = 1;
      ELSE IF ECCFY[M] = 1 AND PMOSX[M+1]='' THEN ELIGREEN1[M] = 0;

*********CHECK FOR ACCESSIONS AND LOSSES;
      IF PMOSX[M]='' AND PMOSX[M+1]~='' THEN NEWBIE[M]=1;
      IF PMOSX[M]~='' AND PMOSX[M+1]='' THEN LOSS[M]=1;
      MOSGRADE[M] = PMOSX[M+1]||GRADE[M+1];
      IF PMOSX[M]~='' AND GRADE[M]~=''
            THEN TRANSITION[M] =
PMOSX[M]||GRADE[M]||'to'||PMOSX[M+1]||GRADE[M+1];
END;

DO P = 1 TO 17;
      IF NEWBIE[P]=. AND PMOSX[P+1]~='' THEN NEWBIE[P]=0;
      IF LOSS[P]=. AND PMOSX[P+1]~='' THEN LOSS[P]=0;
END;

************CLASSIFY ALL MOSs INTO COMBAT(CBT), SERVICE&SUPT(SVC), AND
AVIATION(AVI);
DO Q = 1 TO 9;
      IF OCC[Q] IN ('03' '08' '18') THEN MOSCAT[Q] = 'CBT';
      ELSE IF OCC[Q] IN('60' '61' '62' '63' '64' '65' '66' '70' '72'
'73') THEN MOSCAT[Q] = 'AVI';
      ELSE MOSCAT[Q] = 'SPT';
END;

***********ASSIGN SRB ZONE A, B, OR C, TO MARINES WHO ARE IN A YEAR
***********WITH AN ECC;
DO R = 1 TO 5;
      IF  ECCFY[R+12] NE . THEN DO;
      MONTHSVC[R] = intck('month',ARMED_FORC,lastday[R]+1);
      IF MONTHSVC[R] >= 17 AND MONTHSVC[R] <=72 THEN SRBZ[R] = 'A';
      ELSE IF MONTHSVC[R] > 72 AND MONTHSVC[R] <=120 THEN SRBZ[R] =
'B';
      ELSE IF MONTHSVC[R] > 120 AND MONTHSVC[R] <=168 THEN SRBZ[R] =
'C';
      ELSE SRBZ[R] ='';
      END;
END;

***FIND TOTAL NUMBER OF CONTRACTS MARINE HAS COMPLETED;

ECCTOTAL05 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998, ECCFYTEST1999, ECCFYTEST2000,
             ECCFYTEST2001, ECCFYTEST2002, ECCFYTEST2003,
             ECCFYTEST2004, ECCFYTEST2005);

ECCTOTAL04 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998, ECCFYTEST1999, ECCFYTEST2000,
             ECCFYTEST2001, ECCFYTEST2002, ECCFYTEST2003,
             ECCFYTEST2004);

ECCTOTAL03 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
              ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
              ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
              ECCFYTEST1998, ECCFYTEST1999, ECCFYTEST2000,
              ECCFYTEST2001, ECCFYTEST2002,
              ECCFYTEST2003);

ECCTOTAL02 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998, ECCFYTEST1999, ECCFYTEST2000,
             ECCFYTEST2001, ECCFYTEST2002);

ECCTOTAL01 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998, ECCFYTEST1999, ECCFYTEST2000,
             ECCFYTEST2001);

ECCTOTAL00 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998, ECCFYTEST1999, ECCFYTEST2000);

ECCTOTAL99 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998, ECCFYTEST1999);

ECCTOTAL98 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995,ECCFYTEST1996, ECCFYTEST1997,
             ECCFYTEST1998);

ECCTOTAL97 = SUM(ECCFYTEST1989, ECCFYTEST1990, ECCFYTEST1991,
             ECCFYTEST1992, ECCFYTEST1993, ECCFYTEST1994,
             ECCFYTEST1995, ECCFYTEST1996, ECCFYTEST1997);

DROP  lastday1988-lastday2005 K L M P Q R ECCFYTEST1989-ECCFYTEST2005;
      RUN;
```

Below is the code used for extracting subsets of data for classification trees. The code shown is used for the first-term population only. SRB data is merged with the LDS in this code.

```
*THIS SAS CODE MAKES DATA SETS FOR CLASSIFICATION TREES IN S-PLUS;
*****THE DATA SET MADE HERE IS FOR THE FIRST TERM REENLISTMENT
*****POPULATION FY2004;
*****IT ALSO MERGES THE SRB MULTIPLE OFFERED WITH THE APPROPRIATE
*****GROUPS OF MARINES;
*****CODE FOR SUBSEQUENT TERM POPULATION IS THE SAME EXCEPT ECCTOTAL
*****VARIABLE > 0 INSTEAD OF ECCTOTAL = 0;

LIBNAME CON 'Z:\C';

***CREATE A DATA SET OF ALL ACTIVE DUTY MARINES FROM 1998 TO 2005**;
***YEARS OF DATA NOT USED IN THESIS CAN BE USED IN FUTURE WORK FOR
***MODELING AND VALIDATION;
***TREES WILL BE MADE FOR FIRST TERM AND SUBSEQUENT TERM REENLISTMENT
***MODELING;

DATA CON.JOINEM4;
      SET Demo.joinem3 (DROP = ZONE2001 ZONE2002 ZONE2003 ZONE2004
ZONE2005);
      IF PMOS1998 ~='' OR PMOS1999 ~='' OR PMOS2000 ~='' OR  PMOS2001
~='' OR  PMOS2002 ~='' OR
             PMOS2003 ~='' OR PMOS2004 ~='' OR PMOS2005 ~='';

RUN;

***Bring in data from SRB messages showing which MOSs are offered what
bonus multiple;
***for each year from 2001 - 2005;

PROC IMPORT OUT = CON.SRB01
                  DATAFILE = "Z:\C\SRB01.XLS"
                  DBMS=EXCEL REPLACE;

PROC IMPORT OUT = CON.SRB02
                  DATAFILE = "Z:\C\SRB02.XLS"
                  DBMS=EXCEL REPLACE;

RUN;PROC IMPORT OUT = CON.SRB03
                  DATAFILE = "Z:\C\SRB03.XLS"
                  DBMS=EXCEL REPLACE;

RUN;PROC IMPORT OUT = CON.SRB04
                  DATAFILE = "Z:\C\SRB04.XLS"
                  DBMS=EXCEL REPLACE;

PROC IMPORT OUT = CON.SRB05
                  DATAFILE = "Z:\C\SRB05.XLS"
                  DBMS=EXCEL REPLACE;


RUN;
```

```
*****SORT JOINEM4 BY MOS OF YR PRIOR TO ECC (DUE TO LONGITUDINAL
*****DATASET CONSIDERATIONS) AND ZONE AT SAME TIME;
*****SORT SRB DATA BY MOS, ZONE AT "START" OF ECC YR.  THIS IS THE BEST
*****WAY WE CAN APPROXIMATE ZONE;
*****NEED TO RENAME ONE OF THE MERGING PMOS VAR'S SO THAT THEY MATCH,
*****SINCE WE ARE REALLY MERGING BY YEAR N FROM JOINEM AND N+1 FROM SRB
*****DATA. RENAME THE SRB DATA VARIABLE IN THE EXCEL FILE TO MATCH
*****LDS****;

PROC SORT DATA = CON.joinem4;
      BY PMOS2000 SRBZONE2001;
RUN;

PROC SORT DATA = CON.SRB01;
      BY PMOS2000 SRBZONE2001;
RUN;

DATA MERGE01;
      MERGE CON.JOINEM4 CON.SRB01;
      BY PMOS2000 SRBZONE2001;

RUN;

PROC SORT DATA = MERGE01;
      BY PMOS2001 SRBZONE2002;
RUN;

PROC SORT DATA = CON.SRB02;
      BY PMOS2001 SRBZONE2002;
RUN;

DATA MERGE02;
      MERGE MERGE01 CON.SRB02;
      BY PMOS2001 SRBZONE2002;

RUN;

PROC SORT DATA = MERGE02;
      BY PMOS2002 SRBZONE2003;
RUN;

PROC SORT DATA = CON.SRB03;
      BY PMOS2002 SRBZONE2003;
RUN;


DATA MERGE03;
      MERGE MERGE02 CON.SRB03;
      BY PMOS2002 SRBZONE2003;

RUN;

PROC SORT DATA = MERGE03;
      BY PMOS2003 SRBZONE2004;
RUN;
```

```
PROC SORT DATA = CON.SRB04;
      BY PMOS2003 SRBZONE2004;
RUN;

DATA MERGE04;
      MERGE MERGE03 CON.SRB04;
      BY PMOS2003 SRBZONE2004;

RUN;

PROC SORT DATA = MERGE04;
      BY PMOS2004 SRBZONE2005;
RUN;

PROC SORT DATA = CON.SRB05;
      BY PMOS2004 SRBZONE2005;
RUN;
*MERGE05 DATA SET IS SAVED FOR FUTURE USE - ALL MERGES OF SRB DATA WITH
DEMOGRAPHIC DATA;
*ARE COMPLETED IN THIS DATA SET;

DATA CON.MERGE05;
      MERGE MERGE04 CON.SRB05;
      BY PMOS2004 SRBZONE2005;

RUN;

* Input the data and create some new variables.  Only keep the
variables
  used in the modeling;

DATA CON.temp1;
      SET CON.MERGE05(keep= eligreen12001 eligreen12002 eligreen12003
eligreen12004 eligreen12005 sex race mult2001 mult2002 mult2003
mult2004 mult2005 afqt_score ethnic_gro SRBZONE2001 SRBZONE2002
SRBZONE2003 SRBZONE2004 SRBZONE2005 ECCFY1997 ECCTOTAL97 moscat1997
mosgrade1997 pmos1997 occ1997 grade1997 mar1997 yos1997 dep1997
ECCFY1998 ECCTOTAL98 moscat1998 mosgrade1998 pmos1998 occ1998 grade1998
mar1998 yos1998 dep1998 ECCFY1999 ECCTOTAL99 moscat1999mosgrade1999
pmos1999 occ1999 grade1999 mar1999 yos1999 dep1999 ECCFY2000 ECCTOTAL00
moscat2000  mosgrade2000 pmos2000 occ2000 grade2000 mar2000 yos2000
dep2000 ECCFY2001 ECCTOTAL01 moscat2001 mosgrade2001 pmos2001 occ2001
grade2001 mar2001 yos2001 dep2001 ECCFY2002 ECCTOTAL02 moscat2002
mosgrade2002 pmos2002 occ2002 grade2002 mar2002 yos2002 dep2002
ECCFY2003 ECCTOTAL03 moscat2003 mosgrade2003 pmos2003 occ2003 grade2003
mar2003 yos2003 dep2003 ECCFY2004 ECCTOTAL04 moscat2004mosgrade2004
pmos2004 occ2004 grade2004 mar2004 yos2004 dep2004 ECCFY2005 ECCTOTAL05
moscat2005  mosgrade2005 pmos2005 occ2005 grade2005 mar2005 yos2005
dep2005);

            if mult2001=. then mult2001=0;
            if mult2002=. then mult2002=0;
            if mult2003=. then mult2003=0;
            if mult2004=. then mult2004=0;
            if mult2005=. then mult2005=0;
            ***RACE & ETHNIC CODES;
            *BLACK ==> ETHNIC = 1;
            *HISPANIC ==> ETHNIC = 2;
            *ASIAN ==> ETHNIC = 3;
            *WHITE ==> ETHNIC = 4;
            *PACIFIC ISLANDER ==> ETHNIC = 5;
            *OTHER ==> ETHNIC = 6;
            ETHNIC = '6';
            IF ETHNIC_GRO = 'A' OR (RACE = 'C' AND ETHNIC_GRO = 'Z')
                  THEN ETHNIC = '1';
            IF ETHNIC_GRO IN ('1' '4' '6' '9' 'S')THEN ETHNIC = '2';
            IF ETHNIC_GRO IN ('3' '5' 'G' 'J' 'K' 'V') OR (RACE = 'B'
                  AND ETHNIC_GRO = 'Z') THEN ETHNIC = '3';
            IF ETHNIC_GRO = 'P' OR (RACE = 'E' AND ETHNIC_GRO = 'Z')
                  THEN ETHNIC = '4';
            IF ETHNIC_GRO IN ('E' 'H' 'L' 'Q' 'W') OR (RACE = 'D' AND
                  ETHNIC_GRO = 'Z') THEN ETHNIC = '5';

RUN;

* NOW, CONSTUCT ANALYTIC DATASET IN WHICH THERE IS ONE VARIABLE EACH
FOR   MARITAL STATUS, DEPENDENTS, SRB ELIGIBILITY, GRADE, PMOS AND
OCCFIELD ALL APPROPIATELY LAGGED WITH RESPECT TO THE ELIGREEN1XXXX
VARIABLE;

* WHAT THIS CODE DOES IS TO ASSIGN THE VALUE OF A VARIABLE FOR THE
EARLIEST NON-MISSING ELIGREEN1200X VARIABLE TO THE NEW LAGGED VARIABLE.
IF THE VALUE OF THE VARIABLE IS MISSING FOR THAT YEAR, IT GETS THE
LATEST NON-MISSING VALUE;

DATA CON.FTAP04Tree;
      SET CON.temp1;
      IF (PMOS2003~='' AND ECCFY2004=1 AND ECCTOTAL03 = 0);

      IF (eligreen12004~=. and mar2003~="") THEN marstat=mar2003;
            ELSE IF (eligreen12004~=. and mar2002~="") THEN
                  marstat=mar2002;
            ELSE IF (eligreen12004~=. and mar2001~="") THEN
                  marstat=mar2001;
            ELSE IF (eligreen12004~=. and mar2000~="") THEN
                  marstat=mar2000;

      IF (eligreen12004~=.  and dep2003~=.) THEN depstat=dep2003;
            ELSE IF (eligreen12004~=. and dep2002~=.) THEN
                  depstat=dep2002;
            ELSE IF (eligreen12004~=. and dep2001~=.) THEN
                  depstat=dep2001;
            ELSE IF (eligreen12004~=. and dep2000~=.) THEN
                  depstat=dep2000;
      IF (eligreen12004~=. and grade2003~="") THEN grade=grade2003;
            ELSE IF (eligreen12004~=. and grade2002~="") THEN
                  grade=grade2002;
            ELSE IF (eligreen12004~=. and grade2001~="") THEN
                  grade=grade2001;
            ELSE IF (eligreen12004~=. and grade2000~="") THEN
                  grade=grade2000;

      IF (eligreen12004~=. and pmos2003~="") THEN pmos=pmos2003;
            ELSE IF (eligreen12004~=. and pmos2002~="") THEN
                  pmos=pmos2002;
            ELSE IF (eligreen12004~=. and pmos2001~="") THEN
                  pmos=pmos2001;
            ELSE IF (eligreen12004~=. and pmos2000~="") THEN
                  pmos=pmos2000;

      IF (eligreen12004~=. and occ2003~="") THEN occfield=occ2003;
            ELSE IF (eligreen12004~=. and occ2002~="") THEN
                  occfield=occ2002;
            ELSE IF (eligreen12004~=. and occ2001~="") THEN
                  occfield=occ2001;
            ELSE IF (eligreen12004~=. and occ2000~="") THEN
                  occfield=occ2000;

      IF (eligreen12004~=. and yos2003~="") THEN yos=yos2003;
            ELSE IF (eligreen12004~=. and yos2002~="") THEN
                  yos=yos2002;
            ELSE IF (eligreen12004~=. and yos2001~="") THEN
                  yos=yos2001;
            ELSE IF (eligreen12004~=. and yos2000~="") THEN
                  yos=yos2000;

      IF (eligreen12004~=. and moscat2003~="") THEN moscat=moscat2003;
            ELSE IF (eligreen12004~=. and moscat2002~="") THEN
                  moscat=moscat2002;
            ELSE IF (eligreen12004~=. and moscat2001~="") THEN
                  moscat=moscat2001;
            ELSE IF (eligreen12004~=. and moscat2000~="") THEN
                  moscat=moscat2000;

      IF eligreen12004~=. THEN SRBelig=mult2004;

RUN;
```

This section contains an example of the SAS code used to develop the logistic regression model.  The example shown is for the first-term population.

```
LIBNAME CON 'Z:\C';
LIBNAME LR 'C:\Documents and Settings\dgconats\My
Documents\Thesis\LRData';

* BRING IN FIRST TERM MARINES WHO HAVE REENLISTMENT DECISION TO MAKE IN
FY04 OR FY05;

DATA LR.FtapINT2005;
      SET LR.temp1;
      IF (PMOS2004~='' AND ECCFY2005=1 AND ECCTOTAL04 = 0
            AND YOS2004 >=2 AND YOS2004 <=6 AND GRADE2004 NOT IN ('E7'
            'E8' 'E9'))
      OR (PMOS2003~='' AND ECCFY2004=1 AND ECCTOTAL03 = 0
            AND YOS2003 >=2 AND YOS2003 <=6 AND GRADE2003 NOT IN ('E7'
            'E8' 'E9'));

RUN;

* Now, constuct analytic dataset in which there is one variable each
for marital status, dependents, SRB eligibility, grade, PMOS and
occfield all appropiately lagged with respect to the ELIGREEN1XXXX
variable;

* What this code does is to assign the value of a variable for the
earliest non-missing ELIGREEN1200X variable to the new lagged variable.
If the value of the variable is missing for that year, it gets the
latest non-missing value;

DATA LR.FT05LOOKBACK;
      SET LR.FtapINT2005;

      IF (eligreen12005~=. and mar2004~="") THEN marstat=mar2004;
            ELSE IF (eligreen12005~=. and mar2003~="") THEN
                  marstat=mar2003;
            ELSE IF (eligreen12005~=. and mar2002~="") THEN
                  marstat=mar2002;
            ELSE IF (eligreen12005~=. and mar2001~="") THEN
                  marstat=mar2001;
      IF (eligreen12004~=. and mar2003~="") THEN marstat=mar2003;
            ELSE IF (eligreen12004~=. and mar2002~="") THEN
                  marstat=mar2002;
            ELSE IF (eligreen12004~=. and mar2001~="") THEN
                  marstat=mar2001;
            ELSE IF (eligreen12004~=. and mar2000~="") THEN
                  marstat=mar2000;

      IF (eligreen12005~=.  and dep2004~=.) THEN depstat=dep2004;
            ELSE IF (eligreen12005~=. and dep2003~=.) THEN
                  depstat=dep2003;
            ELSE IF (eligreen12005~=. and dep2002~=.) THEN
                  depstat=dep2002;
            ELSE IF (eligreen12005~=. and dep2001~=.) THEN
                  depstat=dep2001;
```

```
IF (eligreen12004~=. and dep2003~=.) THEN depstat=dep2003;
      ELSE IF (eligreen12004~=. and dep2002~=.) THEN
            depstat=dep2002;
      ELSE IF (eligreen12004~=. and dep2001~=.) THEN
            depstat=dep2001;
      ELSE IF (eligreen12004~=. and dep2000~=.) THEN
            depstat=dep2000;

IF (eligreen12005~=. and grade2004~="") THEN grade=grade2004;
      ELSE IF (eligreen12005~=. and grade2003~="") THEN
            grade=grade2003;
      ELSE IF (eligreen12005~=. and grade2002~="") THEN
            grade=grade2002;
      ELSE IF (eligreen12005~=. and grade2001~="") THEN
            grade=grade2001;
IF (eligreen12004~=. and grade2003~="") THEN grade=grade2003;
      ELSE IF (eligreen12004~=. and grade2002~="") THEN
            grade=grade2002;
      ELSE IF (eligreen12004~=. and grade2001~="") THEN
            grade=grade2001;
      ELSE IF (eligreen12004~=. and grade2000~="") THEN
            grade=grade2000;

IF (eligreen12005~=. and pmos2004~="") THEN pmos=pmos2004;
      ELSE IF (eligreen12005~=. and pmos2003~="") THEN
            pmos=pmos2003;
      ELSE IF (eligreen12005~=. and pmos2002~="") THEN
            pmos=pmos2002;
      ELSE IF (eligreen12005~=. and pmos2001~="") THEN
            pmos=pmos2001;
IF (eligreen12004~=. and pmos2003~="") THEN pmos=pmos2003;
      ELSE IF (eligreen12004~=. and pmos2002~="") THEN
            pmos=pmos2002;
      ELSE IF (eligreen12004~=. and pmos2001~="") THEN
            pmos=pmos2001;
      ELSE IF (eligreen12004~=. and pmos2000~="") THEN
            pmos=pmos2000;

IF (eligreen12005~=. and occ2004~="") THEN occfield=occ2004;
      ELSE IF (eligreen12005~=. and occ2003~="") THEN
            occfield=occ2003;
      ELSE IF (eligreen12005~=. and occ2002~="") THEN
            occfield=occ2002;
      ELSE IF (eligreen12005~=. and occ2001~="") THEN
            occfield=occ2001;
IF (eligreen12004~=. and occ2003~="") THEN occfield=occ2003;
      ELSE IF (eligreen12004~=. and occ2002~="") THEN
            occfield=occ2002;
      ELSE IF (eligreen12004~=. and occ2001~="") THEN
            occfield=occ2001;
      ELSE IF (eligreen12004~=. and occ2000~="") THEN
            occfield=occ2000;

IF (eligreen12005~=. and yos2004~=.) THEN yos=yos2004;
      ELSE IF (eligreen12005~=. and yos2003~=.) THEN yos=yos2003;
      ELSE IF (eligreen12005~=. and yos2002~=.) THEN yos=yos2002;
      ELSE IF (eligreen12005~=. and yos2001~=.) THEN yos=yos2001;
      IF (eligreen12004~=. and yos2003~=.) THEN yos=yos2003;
            ELSE IF (eligreen12004~=. and yos2002~=.) THEN yos=yos2002;
            ELSE IF (eligreen12004~=. and yos2001~=.) THEN yos=yos2001;
            ELSE IF (eligreen12004~=. and yos2000~=.) THEN yos=yos2000;


      IF (eligreen12005~=. and moscat2004~="") THEN moscat=moscat2004;
            ELSE IF (eligreen12005~=. and moscat2003~="") THEN
                  moscat=moscat2003;
            ELSE IF (eligreen12005~=. and moscat2002~="") THEN
                  moscat=moscat2002;
            ELSE IF (eligreen12005~=. and moscat2001~="") THEN
                  moscat=moscat2001;

      IF (eligreen12004~=. and moscat2003~="") THEN moscat=moscat2003;
            ELSE IF (eligreen12004~=. and moscat2002~="") THEN
                  moscat=moscat2002;
            ELSE IF (eligreen12004~=. and moscat2001~="") THEN
                  moscat=moscat2001;
            ELSE IF (eligreen12004~=. and moscat2000~="") THEN
                  moscat=moscat2000;


      IF eligreen12004~=. THEN SRBelig = mult2004;
      IF eligreen12005~=. THEN SRBelig = mult2005;

      IF GRADE IN ('E1' 'E2' 'E3') THEN HIGRADE = 0;
            ELSE HIGRADE = 1;

      IF YOS < 3 THEN YOSI = 0;
            ELSE YOSI = 1;

      IF depstat > 0 then DEPI = 1;
            else DEPI = 0;

      IF ETHNIC = '4' THEN ETHI = 1;
            ELSE ETHI = 0;

      IF AFQT_SCORE < 29 THEN AFQTI = '1';
            ELSE IF 29 =< AFQT_SCORE THEN AFQTI = '2';
            ELSE AFQTI = '3';

      IF moscat = 'CBT' then CBTMOS = 1;
            ELSE CBTMOS = 0;

RUN;

* Find mean reenlistment rate by grade for 2004;
PROC SORT DATA = LR.FT05LOOKBACK;
      BY GRADE;
RUN;
```

```
*RATES FOR '04 WILL BE APPLIED TO '05 WHERE NO Phat EXISTS IN LR MODEL;
PROC MEANS DATA = LR.FT05LOOKBACK MEAN;
      TITLE "REENLISTMENT RATE BY GRADE FOR 2004";
      CLASS GRADE;
      VAR ELIGREEN12004;
      WHERE GRADE ~= '';
      OUTPUT OUT = LR.FT05RATEBYGRADE MEAN = AVRATE;
RUN;

DATA LR.FT05LRMODEL;
      MERGE LR.FT05LOOKBACK LR.FT05RATEBYGRADE;
      BY GRADE;
      DROP _TYPE_ _FREQ_;
RUN;


* The logistic regression model to predict the probability an
  individual will reenlist in 2005.  It is created by first estimating
  model parameters from 2004 data where we know whether each individual
  reenlisted.  Then the model is applied to the particular individuals
  up for reenlistment in 2005 and their probabilities are calculated;

* Interaction syntax: var1|var2 ;

PROC LOGISTIC DATA= LR.FT05LRMODEL DESCENDING ;
      CLASS HIGRADE YOSI DEPI AFQTI CBTMOS ETHI moscat sex ETHNIC
      marstat pmos grade occfield;
      MODEL eligreen12004 = YOSI GRADE ETHI/
      LACKFIT;
      OUTPUT OUT = temp8a PREDICTED = phat;
RUN;

DATA temp9;
      SET temp8a;
      IF eligreen12005 ~= .;
      IF phat=. then phat=AVRATE;

PROC SORT data=temp9;
by pmos grade;

* Estimate the number that reenlist by PMOS and save the data set;
PROC MEANS data=temp9 sum noprint;
var phat;
by pmos grade;
output out=estreupsumbypmos sum=estNreup;
proc print data = estreupsumbypmos; run;
* Calculate the actual number that reenlist by PMOS;
PROC MEANS data=temp9 sum noprint;
var eligreen12005;
by pmos grade;
output out=actreupsumbypmos sum=actNreup;
```

```
* Merge then into one data set;
DATA temp10;
      MERGE estreupsumbypmos actreupsumbypmos;
      BY pmos grade;
      diff=estNreup-actNreup;
      sqdiff=(estNreup-actNreup)*(estNreup-actNreup);
      avgdiff=sqdiff/estNreup;
      IF (estNreup<0.5) THEN estNreup=0.0;
      IF (estNreup=0.0 and actNreup>0) then avgdiff=1.0;
      mosgrade=trim(pmos)||trim(grade);

PROC EXPORT DATA= temp10
            OUTFILE= "Z:\Excel files\FTmodelQ.xls"
            DBMS=EXCEL2000 REPLACE;
RUN;


* Print the results to look at estimates and actuals by PMOS;
proc print data=temp10;


*  Calculate a summary statistic to judge how far off the estimate is.
   Here we use a chi-square-like statistic;

PROC MEANS data=temp10 sum;
   var sqdiff avgdiff;

*  Finally, output the data;
DATA LR.ftap_reup_Ns;
      SET temp10 (keep=mosgrade estNreup);
      label estNreup = "Est Nreup";
RUN;
```

# APPENDIX B.     S-PLUS CODE

```
##### CLASSIFICATION TREE FOR FIRST TERM POPULATION FY2004

> ft04 <- tree (as.factor (ELIGREEN12004) ~  moscat + marstat+ depstat
+ grade + yos + ETHNIC + SRBelig + SEX + AFQT.SCORE, data = ftap04tree,
na.action = na.exclude)
> summary (ft04)
Classification tree:
tree(formula = as.factor(ELIGREEN12004) ~ moscat + marstat + depstat +
grade + yos + ETHNIC + SRBelig + SEX + AFQT.SCORE, data = ftap04tree,
na.action = na.exclude)
Number of terminal nodes:  246
Residual mean deviance:  1.12 = 25260 / 22570
Misclassification error rate: 0.2817 = 6426 / 22811
> ft04.cv.m <- cv.tree (ft04, FUN=prune.misclass)
> ft04.cv.m$size[ft04.cv.m$dev == min(ft04.cv.m$dev)]
[1] 11
> ft04.11 <- prune.misclass (ft04, best=11)
> plot(ft04.11)
> plot(ft04.11, type="u")
> text(ft04.11, pretty = 0)

> summary(ft04.11)

Classification tree:
snip.tree(tree = ft04, nodes = c(980., 5., 60., 14., 31., 123., 6.,
491.))
Variables actually used in tree construction:
[1] "grade"       "yos"          "ETHNIC"      "depstat"     "moscat"
"AFQT.SCORE"
Number of terminal nodes:  11
Residual mean deviance:  1.182 = 26950 / 22800
Misclassification error rate: 0.2942 = 6710 / 22811
> #now put pruned tree in .ps format
>   post.tree   (ft04.11,   "FY04   First   Term   Tree   (Pruned   By
Misclassification Rate)",
+ file = "//comfort//conatser-raymond$/Trees/ft04.m.ps")
```

THIS PAGE INTENTIONALLY LEFT BLANK

# INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Marine Corps Representative
   Naval Postgraduate School
   Monterey, California

4. Director, Training and Education
   MCCDC, Code C46
   Quantico, Virginia

5. Director, Marine Corps Research Center
   MCCDC, Code C40RC
   Quantico, Virginia

6. Operations Officer
   Marine Corps Tactical Systems Support Activity
   Camp Pendleton, California

7. Director, Operations Analysis Division
   MCCDC, Code C45
   Quantico, Virginia

8. Director, Manpower Plans and Policy
   Manpower and Reserve Affairs, Code MPP-50
   Quantico, Virginia

9. Professor Ronald D. Fricker, Jr.
   Department of Operations Research
   Naval Postgraduate School
   Monterey, California

10. Professor Samuel E. Buttrey
    Department of Operations Research
    Naval Postgraduate School
    Monterey, California

11. Deputy Under Secretary of Defense for Military Personnel Policy
Office of the Under Secretary of Defense for Personnel and Readiness
Washington, D.C.

12. Director, Officer and Enlisted Personnel Management
Office of the Under Secretary of Defense for Personnel and Readiness
Washington, D.C.

13. Director, Accession Policy
Office of the Under Secretary of Defense for Personnel and Readiness
Washington, D.C.

14. Dr. James Hosek
Director, Forces and Resources Policy Center
National Defense Research Institute
The RAND Corporation
Santa Monica, California